



HAL
open science

Regression with non compactly supported bases

Fabienne Comte, V Genon-Catalot

► **To cite this version:**

Fabienne Comte, V Genon-Catalot. Regression with non compactly supported bases. 2018. hal-01690856v1

HAL Id: hal-01690856

<https://hal.science/hal-01690856v1>

Preprint submitted on 23 Jan 2018 (v1), last revised 18 Oct 2018 (v4)

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.

REGRESSION WITH NON COMPACTLY SUPPORTED BASES

F. COMTE AND V. GENON-CATALOT

ABSTRACT. This paper is about nonparametric regression function estimation, first in the independent setting and in a second stage, in the context of an autoregressive model or of discrete time observation of a diffusion process, both settings corresponding to dependent variables. Our estimator is a one step projection estimator obtained by least-squares contrast minimization. The specificity of our work is to consider a new model selection procedure including a cutoff for the underlying matrix inversion, and to provide theoretical risk bounds that apply to non compactly supported bases, a case which was specifically excluded of all previous results.

January 23, 2018

1. INTRODUCTION

Consider observations $(X_i, Y_i)_{1 \leq i \leq n}$ drawn from the regression model

$$(1) \quad Y_i = b(X_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma_\varepsilon^2, \quad i = 1, \dots, n.$$

The random design variables $(X_i)_{1 \leq i \leq n}$ are real-valued, independent and identically distributed (i.i.d.) with common density denoted by f , the noise variables $(\varepsilon_i)_{1 \leq i \leq n}$ are i.i.d. real-valued and the two sequences are independent. The problem is to estimate the function $b(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ from observations $(X_i, Y_i)_{1 \leq i \leq n}$.

Classical nonparametric estimation strategies are of two types. First, Nadaraya (1964) and Watson (1964) methods rely on quotient estimators of type $\widehat{b} = \widehat{bf}/\widehat{f}$, where \widehat{bf} and \widehat{f} are projection or kernel estimators of bf and f . Those methods are popular, especially in the kernel setting. However, they require the knowledge or the estimation of f (see Efromovich (1999), Tsybakov (2009)) and in the latter case, the choice of two smoothing parameters.

The second method, proposed by Birgé and Massart (1998), Barron *et al.* (1999), and improved by Baraud (2000, 2002), for fixed and random design, is based on a least squares contrast, analogous to the one used for parametric linear regression:

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

minimized over functions t that admit a finite development over some orthonormal A -supported $\mathbb{L}^2(A, dx)$ basis, $A \subset \mathbb{R}$. In other words, this is a projection method where the coefficients of the approximate function in the finite basis play the same role as the regression parameters in the linear model. This strategy solves part of the drawbacks of the first one. Indeed, it provides directly an estimator of b restricted to the set A , a unique

Université Paris Descartes, Laboratoire MAP5, email: fabienne.comte@parisdescartes.fr.

Université Paris Descartes, Laboratoire MAP5, email: valentine.genon-catalot@parisdescartes.fr.

model selection procedure is required and has been proved to realize an adequate squared bias-variance compromise under weak moment conditions on the noise (see Baraud, 2000, 2002). Lastly, there is no quotient to make, and the rate only depends on the regularity index of b , while in the quotient method it also generally depends on the one of f . All these arguments are very favorable to the second strategy.

Noting that the least squares contrast can be rewritten

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

it can be seen that, for a given function t in a finite dimensional linear space included in $\mathbb{L}^2(A, dx)$, three norms must be compared: the integral $\mathbb{L}^2(A, dx)$ -norm, $\|t\|_A^2 = \int_A t^2(x)dx$, associated with the basis, the empirical norm involved in the definition of the contrast, $\|t\|_n^2 = n^{-1} \sum_{i=1}^n t^2(X_i)$, and its expectation, corresponding to a $\mathbb{L}^2(A, f(x)dx)$ -norm, $\|t\|_f^2 = \int_A t^2(x)f(x)dx$. Due to this difficulty, only compactly supported bases have been considered i.e. the set A on which estimation is done is generally assumed to be compact. This allows to assume that f is lower bounded on A , a condition which would not hold on non compact A . Then, if f is upper and lower bounded on A , the $\mathbb{L}^2(A, f(x)dx)$ and the $\mathbb{L}^2(A, dx)$ norms are equivalent and this makes the problem simpler.

Our aim in this work is to obtain theoretical results in regression function estimation by a projection method in the case of non compact support A of the basis. Indeed, several bases, such as the Laguerre ($A = \mathbb{R}^+$) or the Hermite ($A = \mathbb{R}$) basis, are not compactly supported. Nonparametric density estimation by a projection method on these bases has been the subject of several recent contributions (see *e.g.* Comte *et al.* 2015, Comte and Genon-Catalot, 2015, 2017, Belomestny *et al.* 2016), showing that these bases are convenient and easy to handle. In this paper, we propose a new definition of the model selection procedure for regression function estimation on a set A whether compact or not and prove that it reaches a bias-variance tradeoff in a way that generalizes part of Baraud's (2002) theorems to the non compact case. We also extend the method to dependent models, namely autoregressive models in geometric β -mixing framework (extension of Baraud *et al.* (2001a)) and discretely observed diffusion models (extension of Comte *et al.* (2007)). The plan of the paper is the following. We first define the estimator and check that the most elementary risk bound holds without any basis or support constraint. Then we propose a model selection strategy on a random collection of models taking into account a possible inversion problem of the matrix allowing a unique definition of the estimator. A risk bound for the adaptive estimator is provided: it generalizes existing results to non compactly supported bases. In Section 3, we show how to recover previous results for compactly supported bases, while the Hermite and Laguerre bases are considered in Section 4. Lastly, Section 5 is devoted to dependent contexts: discrete time autoregressive and discrete observation of a diffusion process. Most proofs are gathered in Section 7 while Section 8 gives theoretical tools used along the proofs. Section 6 gives some concluding remarks. An appendix is devoted to numerical illustrations.

2. REGRESSION MODEL

Recall that f denotes the density of X_1 . In the following, $\|\cdot\|_{2,p}$ denotes the euclidean norm in \mathbb{R}^p . For $A \subset \mathbb{R}$, $\|\cdot\|_A$ denotes the integral norm in $\mathbb{L}^2(A, dx)$, $\|\cdot\|_f$ the integral norm in $\mathbb{L}^2(A, f(x)dx)$ and $\|\cdot\|_\infty$ the supremum norm on A .

2.1. Definition of the projection estimator. Consider model (1). Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 0, \dots, m-1)$ be an orthonormal system of A -supported functions belonging to $\mathbb{L}^2(A, dx)$. Define $S_m = \text{span}(\varphi_0, \dots, \varphi_{m-1})$, the linear space spanned by $(\varphi_0, \dots, \varphi_{m-1})$. Note that the φ_j 's may depend on m but for simplicity, we omit this in the notation. We assume that for all j , $\int \varphi_j^2(x) f(x) dx < +\infty$ so that $S_m \subset \mathbb{L}^2(A, f(x)dx)$ and define a projection estimator of the regression function b on A , by

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

where $\gamma_n(t)$ is defined in (2). Clearly, $\gamma_n(t) = n^{-1} \sum_{i=1}^n [Y_i - t(X_i)]^2 - n^{-1} \sum_{i=1}^n Y_i^2$, so that we recognize a classical least squares contrast. For functions s, t , we set

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

and write

$$\langle \vec{u}, t \rangle_n = \frac{1}{n} \sum_{i=1}^n u_i t(X_i)$$

when \vec{u} is the vector $(u_1, \dots, u_n)'$, \vec{u}' denotes the transpose of \vec{u} and t is a function. We introduce the classical matrices

$$\hat{\Phi}_m = (\varphi_j(X_i))_{1 \leq i \leq n, 0 \leq j \leq m-1},$$

and

$$(3) \quad \hat{\Psi}_m = (\langle \varphi_j, \varphi_k \rangle_n)_{0 \leq j, k \leq m-1} = \frac{1}{n} \hat{\Phi}_m' \hat{\Phi}_m, \quad \Psi_m = \left(\int \varphi_j(x) \varphi_k(x) f(x) dx \right)_{0 \leq j, k \leq m-1} = \mathbb{E}(\hat{\Psi}_m).$$

We set $\vec{Y} = (Y_1, \dots, Y_n)'$, and define $\vec{a}^{(m)} = (\hat{a}_0^{(m)}, \dots, \hat{a}_{m-1}^{(m)})'$ as the m -dimensional vector such that $\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j^{(m)} \varphi_j$. Classical computations give, assuming that $\hat{\Psi}_m$ is invertible, that

$$(4) \quad \hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j^{(m)} \varphi_j, \quad \text{with} \quad \vec{a}^{(m)} = (\hat{\Phi}_m' \hat{\Phi}_m)^{-1} \hat{\Phi}_m' \vec{Y} = \frac{1}{n} \hat{\Psi}_m^{-1} \hat{\Phi}_m' \vec{Y}.$$

2.2. Risk bound on a fixed space. We now evaluate the risk of the estimator, without any constraint on the basis support.

Proposition 2.1. *Let $(X_i, Y_i)_{1 \leq i \leq n}$ be observations drawn from model (1) and denote by $b_A = b \mathbf{1}_A$. Assume that $b_A \in \mathbb{L}^2(A, f(x)dx)$ and that $\hat{\Psi}_m$ is invertible. Consider the least squares estimator \hat{b}_m of b , given by (4). Then*

$$(5) \quad \mathbb{E}[\|\hat{b}_m - b_A\|_n^2] \leq \inf_{t \in S_m} \left[\int (b_A - t)^2(x) f(x) dx \right] + \sigma_\varepsilon^2 \frac{m}{n},$$

where f denotes the common density of the X_i 's.

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 $\{(t(X_1), \dots, t(X_n))' : t \in S_m\}$ 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,

$$(6) \quad \|\hat{b}_m - b_A\|_n^2 = \|\Pi_m b - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2 = \inf_{t \in S_m} \|t - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2$$

By taking expectation, we obtain

$$(7) \quad \mathbb{E}[\|\hat{b}_m - b_A\|_n^2] \leq \inf_{t \in S_m} \int (t - b_A)^2(x) f(x) dx + \mathbb{E}[\|\hat{b}_m - \Pi_m b\|_n^2].$$

Now we have:

Lemma 2.1. *Under the assumptions of Proposition 2.1,*

$$\mathbb{E}[\|\hat{b}_m - \Pi_m b\|_n^2] = \sigma_\varepsilon^2 \frac{m}{n}.$$

The result of the previous Lemma can be plugged in (7), thus we obtain Proposition 2.1. \square

The result above is general in the sense that it holds for any basis support, whether compact or not. Moreover, let us stress that Lemma 2.1 is an equality: the variance term is **exactly** equal to $\sigma_\varepsilon^2 m/n$, and this does not depend on the basis.

It is not obvious from (5) that the bias term is small when m is large. The following Lemma gives assumptions ensuring that the bias tends to zero when m grows to infinity.

Lemma 2.2. *Assume that $(\varphi_j)_{j \geq 0}$ is an orthonormal basis of $\mathbb{L}^2(A, dx)$ such that, for all $j \geq 0$, $\int \varphi_j^2(x) f(x) dx < +\infty$, that f is bounded on A and that for all $x \in A$, $f(x) > 0$.*

Then $\inf_{t \in S_m} [\int (b_A - t)^2(x) f(x) dx]$ tends to 0 when m tends to infinity.

The proof of Lemma 2.2 is elementary and relies on the following remarks. Note that $\int (b_A - t)^2(x) f(x) dx = \|b_A - t\|_f^2 = \|b_A \sqrt{f} - t \sqrt{f}\|_A^2$. Under the assumptions of Lemma 2.2, the system $\phi_j = \varphi_j \sqrt{f}$, $j \geq 0$ is a complete family of $\mathbb{L}^2(A, dx)$. Indeed, if $g \in \mathbb{L}^2(A, dx)$, $\int g \phi_j = 0$, $\forall j \geq 0$ means that $\int \varphi_j(g \sqrt{f}) = 0 \forall j \geq 0$ and implies $g = 0$ using our assumptions.

As a consequence, the bias is getting small when m grows, but the variance increases: a compromise has to be found, if m can be chosen.

Remark 2.1. • *Note that the result of Proposition 2.1 holds for any sequence $(X_i)_{1 \leq i \leq n}$ provided that it is independent of $(\varepsilon_i)_{1 \leq i \leq n}$ with i.i.d. centered ε_i .*

• *Heteroskedastic regression. Consider the model*

$$(8) \quad Y_i = b(X_i) + \sigma(X_i) \varepsilon_i, \quad \text{Var}(\varepsilon_1) = \mathbb{E}(\varepsilon_1^2) = 1$$

and the same contrast. Thus the estimator on S_m is still given by (4). Then we can prove that under the assumptions of Proposition 2.1,

$$(9) \quad \begin{aligned} \mathbb{E}[\|\hat{b}_m - b_A\|_n^2] &\leq \inf_{t \in S_m} \left[\int (b_A - t)^2(x) f(x) dx \right] + \mathbb{E} \left[\max_{1 \leq i \leq n} \sigma^2(X_i) \right] \frac{m}{n} \\ &\leq \inf_{t \in S_m} \left[\int (b_A - t)^2(x) f(x) dx \right] + c^2 \frac{m}{n}, \end{aligned}$$

if for all x , $\sigma^2(x) \leq c^2$.

2.3. Adaptive procedure and risk bound. We consider now a collection of nested spaces S_m (that is $S_m \subset S_{m'}$ for $m \leq m'$) such that, for each m , the basis $(\varphi_0, \dots, \varphi_{m-1})$ of S_m satisfies

$$(10) \quad \left\| \sum_{j=0}^{m-1} \varphi_j^2 \right\|_{\infty} \leq c_{\varphi}^2 m \quad \text{for } c_{\varphi}^2 > 0 \quad \text{a constant.}$$

For M a matrix, we denote by $\|M\|_{\text{op}}$ the operator norm defined as the square root of the largest eigenvalue of MM^t . If M is symmetric positive definite, it coincides with its largest eigenvalue.

To select the most relevant space S_m , we proceed by choosing

$$(11) \quad \hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left\{ -\|\hat{b}_m\|_n^2 + \kappa \sigma_{\varepsilon}^2 \frac{m}{n} \right\}$$

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

$$(12) \quad \widehat{\mathcal{M}}_n = \left\{ m \in \{1, 2, \dots, n\}, m(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq 4\mathfrak{c} \frac{n}{\log(n)} \right\}, \quad \text{with } \mathfrak{c} = \left(6 \wedge \frac{1}{\|f\|_{\infty}} \right) \frac{1}{48 c_{\varphi}^2}.$$

In practice, we set $\hat{b}_m^T = \hat{b}_m$ if $m\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \leq 4\mathfrak{c}n/\log(n)$, and $\hat{b}_m^T = 0$ otherwise. A theoretical counterpart of $\widehat{\mathcal{M}}_n$ is useful:

$$(13) \quad \mathcal{M}_n = \left\{ m \in \{1, 2, \dots, n\}, m(\|\Psi_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{c} \frac{n}{\log(n)} \right\},$$

where \mathfrak{c} is defined in (12).

To justify (11), let us explain how each term is related to the bias or the variance obtained in Proposition 2.1. The squared bias term is equal to $\|b_A - b_m^f\|_f^2 = \|b_A\|_f^2 - \|b_m^f\|_f^2$ where b_m^f is the $\mathbb{L}^2(A, f(x)dx)$ -orthogonal projection of b on S_m . The first term $\|b_A\|_f^2$ is unknown but does not depend on m ; on the other hand, $\|b_m^f\|_f^2 = \mathbb{E}[\|b_m^f\|_n^2]$. Thus, the quantity $-\|\hat{b}_m\|_n^2$ approximates the squared bias, up to an additive constant, while $\sigma_{\varepsilon}^2 m/n$ has the variance order. The procedure aims at performing an automatic bias-variance tradeoff.

Theorem 2.1. *Let $(X_i, Y_i)_{1 \leq i \leq n}$ be observations from model (1). Assume that:*

- for each m , the basis of S_m satisfies (10).
- $\|f\|_{\infty} < +\infty$,
- $\mathbb{E}(\varepsilon_1^6) < +\infty$ and $\mathbb{E}[b^4(X_1)] < +\infty$.

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

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

where C is a numerical constant and C' is a constant depending on $f, b, \sigma_{\varepsilon}$.

Remark 2.2. *The constant \mathfrak{c} in the definition of $\widehat{\mathcal{M}}_n$ depends on $\|f\|_{\infty}$ which is unknown. In practice, this quantity has to be replaced by a rough estimator. Otherwise, we can replace the bound $4\mathfrak{c}n/\log(n)$ by $n/\log^2(n)$ in the definitions of the sets $\widehat{\mathcal{M}}_n, \mathcal{M}_n$.*

The constant σ_ε^2 is also generally unknown, and must be replaced by an estimator. We simply propose to use the residual least-squares estimator:

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{b}_{m^*}(X_i))^2$$

where m^* is an arbitrarily chosen dimension, which must be neither too large, nor too small; for instance $m^* = \lfloor \sqrt{n} \rfloor$. See e.g. Baraud (2000), section 6.

A bound for the more classical \mathbb{L}^2 -risk can be deduced from Theorem 2.1.

Corollary 2.1. *Under the assumptions of Theorem 2.1, there exists a numerical constant κ_0 such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}[\|\hat{b}_{\hat{m}} - b_A\|_f^2] \leq C_1 \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \|b_A - t\|_f^2 + \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C'_1}{n}$$

where C_1 is a numerical constant and C'_1 is a constant depending on f , b , σ_ε .

The key tool for proving Theorem 2.1 is Proposition 2.2 which relies on a matricial Bernstein deviation inequality proved in Tropp (2015). The result encompasses all possible classical bases, whether compactly supported or not.

Proposition 2.2. *Assume that X_1, \dots, X_n are i.i.d. with common density f such that $\|f\|_\infty < \infty$. Let $\widehat{\Psi}_m$ be the $m \times m$ matrix defined by Equation (3). Assume that $(\varphi_j)_{0 \leq j \leq m-1}$ satisfies (10). Then for all $u > 0$*

$$\mathbb{P} \left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u \right] \leq 2m \exp \left(- \frac{nu^2/2}{c_\varphi^2 m (\|f\|_\infty + u/3)} \right).$$

3. CASE OF COMPACT A AND COMPACTLY SUPPORTED BASES

In this section, we show that Theorem 2.1 contains and improves classical results when the bases are regular and compactly supported, a case considered by most authors.

Let us give first examples of such bases; for simplicity, we take $A = [0, 1]$. Classical compactly supported bases are: histograms $\varphi_j(x) = \sqrt{m} \mathbf{1}_{[j/m, (j+1)/m]}(x)$, for $j = 0, \dots, m-1$; piecewise polynomials with degree r (rescaled Legendre basis up to degree r on each subinterval $[j/m_r, (j+1)/m_r[$, with $m = (r+1)m_r$); compactly supported wavelets; trigonometric basis with odd dimension m , $\varphi_0(x) = \mathbf{1}_{[0,1]}(x)$ and $\varphi_{2j-1}(x) = \sqrt{2} \cos(2\pi jx) \mathbf{1}_{[0,1]}(x)$, and $\varphi_{2j}(x) = \sqrt{2} \sin(2\pi jx) \mathbf{1}_{[0,1]}(x)$ for $j = 1, \dots, (m-1)/2$.

All these collections satisfy (10) with $c_\varphi^2 = 1$ for histograms and trigonometric basis, $c_\varphi^2 = r+1$ for piecewise polynomials with degree r . The trigonometric spaces are nested; for histograms and piecewise polynomials, the models are nested if the subdivisions are dyadic ($m = 2^k$ for increasing values of k).

When the basis has compact support A , one can assume that f is lower bounded on A , by say f_0 . This assumption is commonly used in papers on nonparametric regression.

It implies that, for $\vec{u} = (u_0, \dots, u_{m-1})'$ a vector of \mathbb{R}^m ,

$$(14) \quad \vec{u}' \Psi_m \vec{u} = \int_A \left(\sum_{j=0}^{m-1} u_j \varphi_j(x) \right)^2 f(x) dx \geq f_0 \int_A \left(\sum_{j=0}^{m-1} u_j \varphi_j(x) \right)^2 dx = f_0 \|\vec{u}\|_{2,m}^2.$$

Therefore $\|\Psi_m^{-1}\|_{\text{op}} \leq 1/f_0$ and $\mathcal{M}_n = \{m, m \leq c'n/\log(n)\}$. The unknown matrix Ψ_m no more appears in the definition of \mathcal{M}_n . Therefore, there is no need to consider $\widehat{\mathcal{M}}_n$ and \widehat{m} is chosen in $\mathcal{M}_n = \{m, m \leq c'n/\log(n)\}$. The results of Theorem 2.1 and Corollary 2.1 in this case correspond to case (K1) of Theorem 1.1 (see also inequality (15)) in Baraud (2002, p.132), under similar moment condition on the noise. Note that our constraint $m \leq c'n/\log(n)$ is better than the one imposed in Baraud (the constraint in Baraud (2002) for a non localized basis such as the trigonometric basis is $m \leq c\sqrt{n/\log^3(n)}$ and thus stronger).

Remark 3.1. If $b_A \in \mathbb{L}^2(A, dx)$ and f is upper bounded by f_1 , $\forall t \in S_m$, $\|b_A - t\|_f^2 \leq f_1 \|b_A - t\|_A^2$ and thus

$$(15) \quad \inf_{t \in S_m} \|b_A - t\|_f^2 \leq f_1 \|b_A - b_m\|_A^2$$

where b_m is the $\mathbb{L}^2(dx)$ -orthogonal projection of b_A on S_m . Now, if b_A belongs to a Besov ball $\mathcal{B}_{\alpha, 2, \infty}(A, R)$ (see De Vore and Lorentz (1993), or Baraud (2002, section 2)), then we get that $\inf_{t \in S_m} \|b_A - t\|_f^2 \lesssim m^{-2\alpha}$. So, choosing $m_{\text{opt}} = n^{1/2\alpha+1}$, we recover the optimal classical rate of convergence (namely a rate of order $n^{-2\alpha/(2\alpha+1)}$) for the risk bound of the adaptive estimator.

4. CASE OF NON COMPACT A AND NON COMPACTLY SUPPORTED BASES

In the case of non compact A , we can not assume that f is lower bounded on A , and therefore we can not get rid of the matrix Ψ_m . Our contribution is to take into account and enlight the role of Ψ_m and to introduce a new selection procedure involving $\widehat{\mathcal{M}}_n$.

We illustrate our general result through two concrete examples of non compactly supported bases: the Laguerre basis on $A = \mathbb{R}^+$ and the Hermite basis on $A = \mathbb{R}$. See e.g. Comte and Genon-Catalot (2017) for density estimation by projection using these bases. We assume in this section that $b_A \in \mathbb{L}^2(A, f(x)dx)$ and that f is A -supported, with $f(x) > 0$, $\forall x \in A$, and upper bounded. This means that $\mathbb{E}(b_A^2(X_1)) < +\infty$, which is not a strong assumption. We do not want to assume that $b_A \in \mathbb{L}^2(A, dx)$, which would exclude for instance linear or polynomial functions.

- Laguerre basis, $A = \mathbb{R}^+$. Consider the Laguerre polynomials (L_j) and the Laguerre functions (ℓ_j) given by

$$(16) \quad L_j(x) = \sum_{k=0}^j (-1)^k \binom{j}{k} \frac{x^k}{k!}, \quad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x \geq 0}, \quad j \geq 0.$$

The collection $(\ell_j)_{j \geq 0}$ constitutes a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$, and is such that (see Abramowitz and Stegun (1964)):

$$(17) \quad \forall j \geq 0, \quad \forall x \in \mathbb{R}^+, \quad |\ell_j(x)| \leq \sqrt{2}.$$

Clearly, the collection of models ($S_m = \text{span}\{\ell_0, \dots, \ell_{m-1}\}$) is nested, and (17) implies that this basis satisfies the general assumption (10) with $c_\varphi^2 = 2$. For a function $\theta \in \mathbb{L}^2(\mathbb{R}^+, dx)$, we can develop θ on the Laguerre basis with: $\theta = \sum_{j \geq 0} a_j(\theta) \ell_j$, $a_j(\theta) = \langle \theta, \ell_j \rangle = \int \theta(x) \ell_j(x) dx$.

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

$$(18) \quad H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j} (e^{-x^2}), \quad h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad c_j = (2^j j! \sqrt{\pi})^{-1/2}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. When a function θ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, it can be developed in the Hermite basis $\theta = \sum_{j \geq 0} a_j(\theta) h_j$ where $a_j(\theta) = \int_{\mathbb{R}} \theta(x) h_j(x) dx = \langle \theta, h_j \rangle$. The infinite norm of h_j satisfies (see Abramowitz and Stegun (1964), Szegö (1959) p.242):

$$(19) \quad \|h_j\|_{\infty} \leq \Phi_0, \quad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160,$$

so that the Hermite basis satisfies the general assumption (10) with $c_{\varphi}^2 = \Phi_0^2$. The collection of models is also clearly nested.

Hereafter, we use the notation φ_j to denote ℓ_j in the Laguerre case and h_j in the Hermite case and denote by $S_m = \text{span}(\varphi_0, \varphi_1, \dots, \varphi_{m-1})$ the linear space generated by the m functions $\varphi_0, \dots, \varphi_{m-1}$ and by $f_m = \sum_{j=0}^{m-1} a_j(f) \varphi_j$ the orthogonal projection of f on S_m . Then $a_j(f) = \langle f, \varphi_j \rangle$ will mean the integral of $f \varphi_j$ either on \mathbb{R} or on \mathbb{R}^+ .

As the bases functions are bounded, the terms $\int \varphi_j^2 f$ are finite. Moreover, the assumptions of Lemma 2.2 hold, so that the bias term in Proposition 2.1 tends to zero as m grows to infinity.

The matrices Ψ_m and $\widehat{\Psi}_m$ in these bases have specific properties.

Lemma 4.1. *For all $m \in \mathbb{N}$, Ψ_m is invertible, and for all $m \leq n$, $\widehat{\Psi}_m$ is invertible.*

Proof of Lemma 4.1. Using (14), for all $\vec{u} = (u_0, \dots, u_{m-1})' \in \mathbb{R}^m \setminus \{\vec{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j \varphi_j(x)$, $\vec{u}' \Psi_m \vec{u} = \|t\|_A^2 > 0$. Moreover $\vec{u}' \widehat{\Psi}_m \vec{u} = \|t\|_n^2 \geq 0$. Thus $\|t\|_n = 0 \Rightarrow t(X_i) = 0$ for $i = 1, \dots, n$. As $t(x)w(x)$ is a polynomial with degree $m-1$ where $w(x) = e^x$ in the Laguerre case and $w(x) = e^{x^2/2}$ in the Hermite case, for $m \leq n$, we obtain that $t \equiv 0$. This implies $\vec{u} = \vec{0}$. \square

Proposition 4.1. *Assume that $\inf_{a \leq x \leq b} f(x) > 0$ for some interval $[a, b]$ in the Hermite case and with $0 < a < b$ in the Laguerre case. Then there exists a constant c^* such that, for all m ,*

$$(20) \quad \|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^* m.$$

In the Laguerre and Hermite cases, Inequality (20) clearly implies that $\|\Psi_m^{-1}\|_{\text{op}}$ cannot be uniformly bounded in m contrary to the case of compactly supported bases. This means that the constraint in the definition (13) of \mathcal{M}_n leads to restrictions on the values m that can be considered in the upper risk bound of Theorem 2.1. This is illustrated by the next proposition.

Proposition 4.2. *Consider the Laguerre or the Hermite basis. Assume that $f(x) \geq c/(1+x)^k$ for $x \geq 0$ in the Laguerre case; or $f(x) \geq c/(1+x^2)^k$ for $x \in \mathbb{R}$ in the Hermite case. Then for m large enough, $\|\Psi_m^{-1}\|_{\text{op}} \leq Cm^k$.*

Discussion. • The inequality given in Proposition 4.2 seems to give the precise order of $\|\Psi_m^{-1}\|_{\text{op}}$. We illustrate it for the Laguerre basis. Indeed, for the density $f_k(x) =$

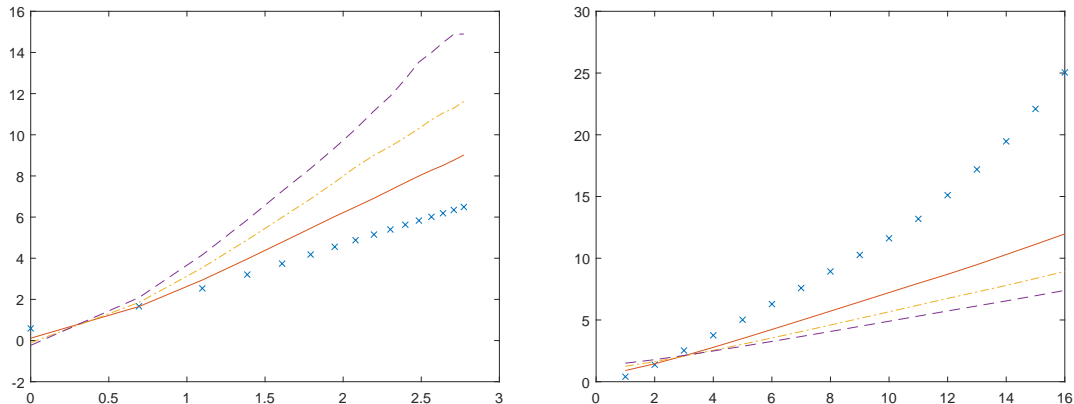


FIGURE 1. Laguerre basis. Left: $\log(m) \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ and density of X given by $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$. Right: $m \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$, and density of X given by $f_k(x) = (k-1)e^{-x/(k-1)} \mathbf{1}_{x \geq 0}$. In both cases: $k = 2$ (blue x marks), $k = 3$ (red solid), $k = 4$ (yellow dashdots) and $k = 5$ (purple dashed).

k	2	3	4	5
\hat{b}_1	2.09	3.16	4.21	5.58
\hat{b}_2	0.68	1.44	2.05	2.67
$\hat{b}_2(k)/\hat{b}_2(2)$	1.00	2.11	3.02	3.92

TABLE 1. Estimated slope regression coefficients, \hat{b}_1 for left curves and \hat{b}_2 for right curves, of Figure 1.

$(k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$, we have computed a Monte-Carlo approximation of $\|\Psi_m^{-1}\|_{\text{op}}$ via 500 samples of size $n = 1000$ and plot in Figure 1 $\log(m) \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ for $m = 1, \dots, 16$. Then we observe that these curves are linear and with slope approximately equal to k (see Table 1). For the density $(k-1)e^{-x/(k-1)} \mathbf{1}_{x \geq 0}$, from the proof of Proposition 4.2, we conjecture that $m \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ is linear with slope proportional to $1/k$; this is confirmed by Figure 1 and the last two lines of Table 1. Figure 1 also shows that the numerical values of $\|\Psi_m^{-1}\|_{\text{op}}$ and thus of $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$ are very quickly increasing and thus few elements are considered in $\widehat{\mathcal{M}}_n$. Nevertheless, the selected \hat{m} among these values provides a very satisfactory estimator of b . The procedure is quick and easy. All this is more detailed in a supplementary material.

• If f is as in Proposition 4.2, the values m in the set \mathcal{M}_n are restricted by the constraint $m^{2k+1} \lesssim n/\log(n)$. This has consequences. Consider the case $A = \mathbb{R}^+$ and the Laguerre basis. Assume that b_A belongs to the Sobolev-Laguerre ball with regularity index s :

$$W^s(R) = \{h \in \mathbb{L}^2(\mathbb{R}^+, dx), \sum_{j \geq 0} a_j(h) j^s \leq R\},$$

with $a_j(h) = \langle h, \ell_j \rangle$, and that f is upper bounded and as in Proposition 4.2. Then, $\inf_{t \in S_m} \|b_A - t\|_f^2 \lesssim m^{-s}$ (see (15) and Remark 3.1 for the analogous computation on a Besov ball). The best compromise between the bias term m^{-s} and the variance term m/n is obtained for $m_{\text{opt}} = n^{1/(s+1)}$. However, the resulting optimal rate of order $n^{-s/(s+1)}$ cannot be reached by the adaptive estimator unless $m_{\text{opt}}^{2k+1} \leq n/\log(n)$. This imposes $s > 2k$. In a Sobolev-Laguerre ball, the index s is linked with regularity properties of functions (see Section 7 of Comte and Genon-Catalot (2015) and Section 7.2 of Belomestny *et al.* (2016)). Thus, the optimal rate cannot be attained by the adaptive estimator unless the function b_A has regularity $s > 2k$.

If b_A is a combination of Γ functions, then the bias term $\inf_{t \in S_m} \|b_A - t\|_f^2$ is much smaller and the rate $\log(n)/n$ can be reached by the adaptive estimator.

In density estimation using projection methods on Laguerre or Hermite bases, the variance term in the risk bound of projection estimators has order \sqrt{m}/n so that the optimal rate on a Sobolev-Laguerre or Sobolev-Hermite ball for the estimators risk is $n^{-2s/(2+1)}$. It seems that, in the regression setting, we cannot have such a gain.

Analogous considerations hold with the Hermite basis.

5. DEPENDENT MODELS.

In this section, we extend the previous results to dependent models, such as autoregressive or diffusion models. The general method is the same as in the proof of Theorem 3.1 in Baraud *et al.* (2001b) for autoregressive models and in the proof of Theorem 1 in Comte *et al.* (2007), both relying in a martingale deviation inequality and a chaining method. The main difficulty here concerns the extension of the deviation inequality stated in Proposition 2.2.

5.1. Mixing deviation inequality. The deviation inequality of Proposition 2.2 can be extended to the mixing case in the specific case of Laguerre and Hermite bases as follows.

Proposition 5.1. *Assume that $(X_i)_i$ is a strictly stationary and geometrically β -mixing process (i.e. the β -mixing coefficients $(\beta_k)_k$ satisfy $\beta_k \leq ce^{-\theta k}$ for some constants $c > 0, \theta > 0$), with marginal density f and that*

- $\mathbb{E}(X_1^{8/3}) < +\infty$ in the Hermite basis,
- $\mathbb{E}(1/X_1^2) < +\infty$ in the Laguerre basis.

For $\widehat{\Psi}_m$ defined by Equation (3), then for all $u > 0$

$$\mathbb{P} \left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u \right] \leq 2m \exp \left(-\frac{nu^2/2}{\mathbf{a}m(1 + \log(n)u)} \right) + \frac{c}{n^4},$$

where \mathbf{a} is a constant depending on the mixing coefficients and the moments of the assumptions.

5.2. Autoregressive model. Let us consider the autoregression model:

$$(21) \quad X_{i+1} = b(X_i) + \varepsilon_{i+1}, \quad (\varepsilon_i)_{i \geq 0} \text{ i.i.d., centered with variance } \sigma_\varepsilon^2.$$

We assume that X_0 is independent of the sequence $(\varepsilon_i)_{i \geq 0}$.

Conditions on $b(\cdot)$ and the noise density ensuring that the model (21) admits a strictly stationary and geometrically β -mixing solution are given in *e.g.* Doukhan (1994) (Th. 7 p.102), and recalled in Baraud *et al.* (2001b), section 5.2.

The contrast and the collection of estimators are then defined by

$$\hat{b}_m = \arg \min_{t \in S_m} \bar{\gamma}_n(t), \quad \text{with } \bar{\gamma}_n(t) = \frac{1}{n} \sum_{i=1}^n t^2(X_i) - 2X_{i+1}t(X_i).$$

The elementary computation of Proposition 2.1 can not be generalized here, but the general strategy for selecting m given by (11) can be extended. The sets $\widehat{\mathcal{M}}_n$, \mathcal{M}_n are now given by

$$(22) \quad \widehat{\mathcal{M}}_n = \left\{ m \in \{1, 2, \dots, n\}, m(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \bar{c} \frac{n}{\log^2(n)} \right\},$$

$$(23) \quad \overline{\mathcal{M}}_n = \left\{ m \in \{1, 2, \dots, n\}, m(\|\Psi_m^{-1}\|_{\text{op}}^2 \vee m) \leq \bar{c}' \frac{n}{\log^2(n)} \right\},$$

with \bar{c}, \bar{c}' constants depending on \mathbf{a} (the constant appearing in the bound of Proposition 5.1). Then we can generalize to the result of Theorem 2.1, thanks to Proposition 5.1.

Theorem 5.1. *Let $(X_i)_{1 \leq i \leq n+1}$ be $n+1$ observations extracted from a strictly stationary and geometrically β -mixing process obtained from model (21), with marginal density f . We consider the Hermite basis if the X_i are real-valued and $\mathbb{E}(X_1^{8/3}) < +\infty$, and the Laguerre basis if the (X_i) are nonnegative and $\mathbb{E}(1/X_1^2) < +\infty$. We also assume that the $(\varepsilon_i)_i$ are i.i.d. centred random variables with $\mathbb{E}(\varepsilon_1^6) < +\infty$. Then, there exists a numerical constant κ_0 such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}[\|\hat{b}_m - b_A\|_n^2] \leq C \inf_{m \in \overline{\mathcal{M}}_n} \left(\inf_{t \in S_m} \|b_A - t\|_f^2 + \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C'}{n}$$

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

5.3. Diffusion processes. Consider now the framework of observations with sampling interval Δ , $(X_{i\Delta})_{1 \leq i \leq n}$, from the diffusion process

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 \sim \eta.$$

Assumptions ensuring that the process $(X_{i\Delta})_{i \geq 0}$ is strictly stationary (with stationary density f) and geometrically β -mixing are given in *e.g.* Comte *et al.* (2007), Section 2.1 p.516. Defining

$$Y_{i\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta}, \quad Z_{i\Delta} = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} \sigma(X_s) dW_s \quad \text{and} \quad R_{i\Delta} = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} \sigma(X_s) dW_s,$$

the approximate regression equation holds:

$$Y_{i\Delta} = b(X_{i\Delta}) + Z_{i\Delta} + R_{i\Delta},$$

where $Z_{i\Delta}$ plays the role of the noise and $R_{i\Delta}$ is an additional residual term to take into account. We set

$$\hat{b}_m = \arg \min_{t \in S_m} \left[\frac{1}{n} \sum_{i=1}^n t^2(X_{i\Delta}) - 2Y_{i\Delta}t(X_{i\Delta}) \right]$$

and obtain an estimator of b in the model.

Assuming that $\Delta = \Delta_n$ tends to zero while $n\Delta_n$ tends to infinity, we can generalize the result of Theorem 1 in Comte *et al.* (2007) into a formulation similar to Theorem 5.1 with $n\Delta$ replacing n .

6. CONCLUDING REMARKS

In this paper, we study nonparametric regression function estimation by a projection method which was first proposed by Birgé and Massart (1998) and Barron *et al.* (1999). Compared with the popular Nadaraya-Watson approach, the projection method has several advantages.

In the Nadaraya-Watson method, one estimates b by a quotient of estimators, namely $\hat{b} = \widehat{bf}/\widehat{f}$. Dividing by \widehat{f} requires a cutoff or a threshold to avoid too small values in the denominator; determining its level is difficult. It is not clear if bandwidth or model selection must be performed separately or simultaneously for the numerator and the denominator. The rate of the final estimator of b corresponds to the worst rate of the two estimators; in particular, it depends on the regularity index of b , but also on the one of f . Therefore, the rate can correspond to the one associated to the regularity index of b , if f is more regular than b , but it is deteriorated if f is less regular than b .

On the other hand, there is no support constraint for this estimation method.

In the projection method used here, the drawbacks listed above do not perturb the estimation except that the unknown function b is estimated in a restricted domain A . Up to now, this set was always assumed to be compact. In the present paper, we show how to eliminate the support constraint by introducing a new selection procedure where the dimension of the projection space is chosen in a random set. The procedure can be applied to non compactly supported bases such as the Laguerre or Hermite bases.

Our method can be readily extended to the case where the Y_i are not observed but subject to multiplicative noise. More precisely, suppose that the observations are $(Z_i, X_i)_{1 \leq i \leq n}$ with

$$Z_i = Y_i U_i, \quad \mathbb{E}(U_i) = 1, \quad \text{and } (Y_i, X_i) \text{ following model (1).}$$

Assume also that the U_i are i.i.d, and the sequences $(\varepsilon_i)_{1 \leq i \leq n}$, $(X_i)_{1 \leq i \leq n}$ and $(U_i)_{1 \leq i \leq n}$ are independent. In this case, if the matrix $\widehat{\Psi}_m$ is invertible, we define

$$\hat{b}_m = \arg \min_{t \in S_m} \left[\|t\|_n^2 - \frac{2}{n} \sum_{i=1}^n Z_i t(X_i) \right].$$

Note that the model can be written $Z_i = b(X_i) + \eta_i$ with $\eta_i = b(X_i)(U_i - 1) + \varepsilon_i U_i$ and is thus of the same type as (8). Thus, we can prove

$$\begin{aligned} \mathbb{E}[\|\hat{b}_m - b\|_n^2] &\leq \int (b_m - b)^2(x) f(x) dx + 2[\text{Var}(U_1) \mathbb{E}(\max_{1 \leq i \leq n} b^2(X_i)) + \sigma_\varepsilon^2 \mathbb{E}(U_1^2)] \frac{m}{n}, \\ &\leq \int (b_m - b)^2(x) f(x) dx + \frac{1}{3} \left[\frac{\|b\|_\infty^2}{2} + \sigma_\varepsilon^2 \right] \frac{m}{n} \end{aligned}$$

if $b(\cdot)$ is bounded.

Note that similar regression strategies have been used in other problems, for instance survival function estimation for interval censored data (see Brunel and Comte (2009)), hazard rate estimation in presence of censoring (see Plancade (2011)): our proposal for classical regression may extend to these contexts, for which it is natural to use \mathbb{R}^+ -supported bases. Indeed, the variables are lifetimes and thus nonnegative, and censoring implies that the right-hand bound of the support is unknown and difficult to estimate; it is thus most convenient that the Laguerre basis does not require to choose it.

7. PROOFS

7.1. Proofs of Section 2.

7.1.1. *Proof of Lemma 2.1.* 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

$$\hat{b}_m(X) = (\hat{b}_m(X_1), \dots, \hat{b}_m(X_n))' = \hat{\Phi}_m \vec{a}^{(m)},$$

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

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

Now, denoting by $\mathbf{P}(X) := \hat{\Phi}_m (\hat{\Phi}_m' \hat{\Phi}_m)^{-1} \hat{\Phi}_m'$, we get

$$(24) \quad \|\hat{b}_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X) \vec{\varepsilon}\|_n^2 = \frac{1}{n} \vec{\varepsilon}' \mathbf{P}(X)' \mathbf{P}(X) \vec{\varepsilon} = \frac{1}{n} \vec{\varepsilon}' \mathbf{P}(X) \vec{\varepsilon}$$

as $\mathbf{P}(X)$ 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

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

Next, we have to compute, using that $\mathbb{P}(X)$ has coefficients depending on the X_i 's only,

$$\mathbb{E}[\vec{\varepsilon}' \mathbf{P}(X) \vec{\varepsilon}] = \sum_{i,j} \mathbb{E}[\varepsilon_i \varepsilon_j \mathbf{P}_{i,j}(X)] = \sigma_\varepsilon^2 \sum_{i=1}^n \mathbb{E}[\mathbf{P}_{i,i}(X)] = \sigma_\varepsilon^2 \mathbb{E}[\text{Tr}(\mathbf{P}(X))],$$

where $\text{Tr}(\cdot)$ is the trace of the matrix. So, we find

$$\text{Tr}(\mathbf{P}(X)) = \text{Tr}((\hat{\Phi}_m' \hat{\Phi}_m)^{-1} \hat{\Phi}_m' \hat{\Phi}_m) = \text{Tr}(\mathbf{I}_m) = m$$

where \mathbf{I}_m is the $m \times m$ identity matrix. Finally, we get

$$\mathbb{E}[\|\hat{b}_m - \Pi_m b\|_n^2] = \sigma_\varepsilon^2 \frac{m}{n}.$$

This is the result of Lemma 2.1. \square

7.1.2. *Proof of Inequality (9).* Let $\vec{\sigma\varepsilon}$ denotes the $n \times 1$ -vector with coordinates $\sigma(X_i) \varepsilon_i$, $i = 1, \dots, n$. Equality (24) now writes

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

as $\mathbf{P}(X)' \mathbf{P}(X) = \mathbf{P}(X)$. Thus, we have to bound

$$\begin{aligned} \mathbb{E}[(\vec{\sigma\varepsilon})' \mathbf{P}(X) (\vec{\sigma\varepsilon})] &= \sum_{i,j} \mathbb{E}[\varepsilon_i \varepsilon_j \sigma(X_i) \sigma(X_j) [\mathbf{P}(X)]_{i,j}] = \sum_{i=1}^n \mathbb{E}[\sigma^2(X_i) [\mathbf{P}(X)]_{i,i}] \\ &\leq \mathbb{E}\left[\max_{1 \leq i \leq n} \sigma^2(X_i) \text{Tr}(\mathbf{P}(X))\right] \leq m \mathbb{E}\left[\max_{1 \leq i \leq n} \sigma^2(X_i)\right], \end{aligned}$$

where $\mathbf{P}(X)$ is defined in the proof of Lemma 2.1. Finally, we obtain (9). \square

7.1.3. *Proof of Proposition 2.2.* To get the announced result, we apply a Bernstein matrix inequality (see Theorem 8.2). Thus we write $\widehat{\Psi}_m$ as a sum of a sequence of independent matrices

$$\widehat{\Psi}_m = \frac{1}{n} \sum_{i=1}^n \mathbf{K}_m(X_i), \quad \mathbf{K}_m(X_i) = (\varphi_j(X_i)\varphi_k(X_i))_{0 \leq j, k \leq m-1}.$$

We put

$$(25) \quad \mathbf{S}_m = \frac{1}{n} \sum_{i=1}^n \mathbf{K}_m(X_i) - \mathbb{E}[\mathbf{K}_m(X_i)].$$

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

We have that

$$\begin{aligned} \|\mathbf{K}_m(X_1) - \mathbb{E}[\mathbf{K}_m(X_1)]\|_{\text{op}} &= \sup_{\|\vec{x}\|_{2,m}=1} \sum_{0 \leq j, k \leq m-1} x_j x_k \left([\mathbf{K}_m(X_1)]_{j,k} - \mathbb{E}[\mathbf{K}_m(X_1)]_{j,k} \right) \\ &= \sup_{\|\vec{x}\|_{2,m}=1} \sum_{0 \leq j, k \leq m-1} x_j x_k (\varphi_j(X_1)\varphi_k(X_1) - \mathbb{E}[\varphi_j(X_1)\varphi_k(X_1)]) \\ &= \sup_{\|\vec{x}\|_{2,m}=1} \left[\left(\sum_{j=0}^{m-1} x_j \varphi_j(X_1) \right)^2 - \left(\sum_{j=0}^{m-1} x_j \mathbb{E}(\varphi_j(X_1)) \right)^2 \right] \leq c_\varphi^2 m. \end{aligned}$$

So we get that

$$(26) \quad \|\mathbf{K}_m(X_1) - \mathbb{E}[\mathbf{K}_m(X_1)]\|_{\text{op}}/n \leq \frac{c_\varphi^2 m}{n}.$$

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

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

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

$$\begin{aligned} \mathbb{E}_1 &:= \mathbb{E} \|(\mathbf{K}_m(X_1) - \mathbb{E}[\mathbf{K}_m(X_1)]) \vec{x}\|^2 = \sum_{j=0}^{m-1} \text{Var} \left[\sum_{k=0}^{m-1} (\varphi_j(X_1)\varphi_k(X_1)) x_k \right] \\ &\leq \sum_{j=0}^{m-1} \mathbb{E} \left(\sum_{k=0}^{m-1} (\varphi_j(X_1)\varphi_k(X_1)) x_k \right)^2 = \sum_{j=0}^{m-1} \int \left(\sum_{k=0}^{m-1} (\varphi_j(u)\varphi_k(u)) x_k \right)^2 f(u) du \end{aligned}$$

Therefore as f is bounded,

$$\mathbb{E}_1 \leq \|f\|_\infty \sum_{j=0}^{m-1} \int \left(\sum_{k=0}^{m-1} (\varphi_j(u)\varphi_k(u)) x_k \right)^2 du \leq \|f\|_\infty c_\varphi^2 m \sum_{k=0}^{m-1} x_k^2 = \|f\|_\infty c_\varphi^2 m.$$

Then we get that $\nu(\mathbf{S}_m) \leq \frac{c_\varphi^2 \|f\|_\infty m}{n}$. Applying Theorem 8.2 gives the result of Proposition 2.2. \square

7.1.4. *Proof of Theorem 2.1.* Let

$$(27) \quad \hat{m}^{\text{th}} = \arg \min_{m \in \mathcal{M}_n} \left\{ -\|\hat{b}_m\|_n^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right\}$$

the theoretical counterpart of \hat{m} , where \mathcal{M}_n is defined by (13). We first prove the following result.

Theorem 7.1. *Under the assumptions of Theorem 2.1, there exists κ_0 such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}[\|\hat{b}_{\hat{m}^{\text{th}}} - b\|_n^2] \leq C \inf_{m \in \mathcal{M}_n} \left(\int_{t \in S_m} \|t - b_A\|_f^2 + \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C'}{n}$$

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

Proof of Theorem 7.1. To begin with, we note that $\gamma_n(\hat{a}_1, \dots, \hat{a}_m) = -\|\hat{b}_m\|_n^2$. Indeed, using formula (4) and $\widehat{\Phi}'_m \widehat{\Phi}_m = n \widehat{\Psi}_m$, we have

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

Consequently, we can write

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{ \gamma_n(\hat{b}_m) + \text{pen}(m) \}, \quad \text{with} \quad \text{pen}(m) = \kappa \sigma_\varepsilon^2 \frac{m}{n}.$$

Now, using the definition of the contrast, we have, for any $m \in \mathcal{M}_n$, and any $b_m \in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}^{\text{th}}}) + \text{pen}(\hat{m}^{\text{th}}) \leq \gamma_n(b_m) + \text{pen}(m)$$

and, with decomposition (6), it yields

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

where $\nu_n(t) = \langle \vec{\varepsilon}, t \rangle_n$. We introduce, for $\|t\|_f^2 = \int t^2(u) f(u) du$, the unit ball

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

and the set

$$(28) \quad \Omega_n = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \frac{1}{2}, \forall t \in \bigcup_{m,m' \in \mathcal{M}_n} (S_m + S_{m'}) \setminus \{0\} \right\}.$$

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

$$(29) \quad \begin{aligned} \|\hat{b}_{\hat{m}^{\text{th}}} - b\|_n^2 \mathbf{1}_{\Omega_n} &\leq \|b_m - b\|_n^2 + \frac{1}{8} \|\hat{b}_{\hat{m}^{\text{th}}} - b_m\|_f^2 \mathbf{1}_{\Omega_n} + 8 \sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} \nu_n^2(t) + \text{pen}(m) - \text{pen}(\hat{m}^{\text{th}}) \\ &\leq \left(1 + \frac{1}{2}\right) \|b_m - b\|_n^2 + \frac{1}{2} \|\hat{b}_{\hat{m}^{\text{th}}} - b\|_n^2 \mathbf{1}_{\Omega_n} + 8 \left(\sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} \nu_n^2(t) - p(m, \hat{m}^{\text{th}}) \right)_+ \\ &\quad + \text{pen}(m) + 8p(m, \hat{m}^{\text{th}}) - \text{pen}(\hat{m}^{\text{th}}). \end{aligned}$$

Here we state the following Lemma:

Lemma 7.1. *Let $(X_i, Y_i)_{1 \leq i \leq n}$ be observations from model (1), \hat{m}^{th} defined by (27) in \mathcal{M}_n defined by (13). Assume that for each m , the basis of S_m satisfies $\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m$ for $c_\varphi^2 > 0$ a constant and that $\mathbb{E}(\varepsilon_1^6) < +\infty$. Then $\nu_n(t) = \langle \bar{\varepsilon}, t \rangle_n$ satisfies*

$$\mathbb{E} \left(\sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} \nu_n^2(t) - p(m, \hat{m}^{\text{th}}) \right)_+ \leq \frac{C}{n}$$

where $p(m, m') = 8\sigma_\varepsilon^2 \max(m, m')/n$.

We see that, for $\kappa \geq \kappa_0 = 32$, we have $8p(m, \hat{m}^{\text{th}}) - \text{pen}(\hat{m}^{\text{th}}) \leq \text{pen}(m)$. Thus, by taking expectation in (29) and applying Lemma 7.1, it comes that, for all m in \mathcal{M}_n and b_m in S_m ,

$$(30) \quad \mathbb{E}[\|\hat{b}_{\hat{m}^{\text{th}}} - b_A\|_n^2 \mathbf{1}_{\Omega_n}] \leq 3\mathbb{E}[\|b_m - b_A\|_n^2] + 2\text{pen}(m) + \frac{16C}{n}.$$

The complement of Ω_n satisfies the following Lemma:

Lemma 7.2. *Assume that X_1, \dots, X_n are i.i.d. with common density f such that $\|f\|_\infty < \infty$. Assume also that the $(\varphi_j)_{0 \leq j \leq m-1}$ are such that $\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m$. Then, for all $m \in \mathcal{M}_n$ (see (13)) and Ω_n defined by (28), $\mathbb{P}(\Omega_n^c) \leq c/n^4$ where c is a positive constant.*

We prove now that $\mathbb{E}[\|\hat{b}_{\hat{m}^{\text{th}}} - b\|_n^2 \mathbf{1}_{\Omega_n^c}] \leq \frac{c}{n}$. Recall that Π_m denotes the orthogonal projection (for the scalar product of \mathbb{R}^n) on the sub-space $\{(t(X_1), \dots, t(X_n))' : t \in S_m\}$ of \mathbb{R}^n . We have $(\hat{b}_m(X_1), \dots, \hat{b}_m(X_n))' = \Pi_m Y$. By using the same notation for the function t and the vector $(t(X_1), \dots, t(X_n))'$, we can see that

$$(31) \quad \|b - \hat{b}_{\hat{m}^{\text{th}}}\|_n^2 = \|b - \Pi_{\hat{m}^{\text{th}}} b\|_n^2 + \|\Pi_{\hat{m}^{\text{th}}} \varepsilon\|_n^2 \leq \|b\|_n^2 + n^{-1} \sum_{k=1}^n \varepsilon_k^2.$$

Thus

$$\begin{aligned} \mathbb{E}[\|b - \hat{b}_{\hat{m}^{\text{th}}}\|_n^2 \mathbf{1}_{\Omega_n^c}] &\leq \mathbb{E}[\|b\|_n^2 \mathbf{1}_{\Omega_n^c}] + \frac{1}{n} \sum_{k=1}^n \mathbb{E}[\varepsilon_k^2 \mathbf{1}_{\Omega_n^c}] \\ &\leq \left(\sqrt{\mathbb{E}[b^4(X_1)]} + \sqrt{\mathbb{E}[\varepsilon_1^4]} \right) \sqrt{\mathbb{P}(\Omega_n^c)}. \end{aligned}$$

Under the assumptions of theorem 7.1, we deduce that

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

This result, together with (30) ends the proof of Theorem 7.1. \square

Proof of Lemma 7.1. We can not apply Talagrand's Inequality to the process ν_n itself, unless we add an assumption imposing that the noise is bounded. This is why we decompose the variables ε_i as follows:

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

Then we have

$$\nu_n(t) = \nu_{n,1}(t) + \nu_{n,2}(t), \quad \nu_{n,1}(t) = \langle \eta, t \rangle_n, \quad \nu_{n,2}(t) = \langle \xi, t \rangle_n,$$

and

$$(32) \quad \left(\sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} \nu_n^2(t) - p(m, \hat{m}^{\text{th}}) \right)_+ \leq \left(\sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} 2\nu_{n,1}^2(t) - p(m, \hat{m}^{\text{th}}) \right)_+ + 2 \sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} \nu_{n,2}^2(t).$$

We successively bound the two terms.

Let $(\bar{\varphi}_j)_{j \in \{1, \dots, \max(m, m')\}}$ be an orthonormal basis of $S_m + S_{m'}$ for the weighted scalar product $\langle \cdot, \cdot \rangle_f$.

It is easy to see that:

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in B_{m', m}^f(0,1)} \nu_{n,1}^2(t) \right] &\leq \sum_{j \leq \max(m, m')} \frac{1}{n} \text{Var} \left(\eta_1 \bar{\varphi}_j(X_1) \right) \leq \sum_{j \leq \max(m, m')} \frac{1}{n} \mathbb{E} \left[\left(\eta_1 \bar{\varphi}_j(X_1) \right)^2 \right] \\ &\leq \frac{1}{n} \mathbb{E}[\varepsilon_1^2] \sum_{j \leq \max(m, m')} \mathbb{E}[\bar{\varphi}_j^2(X_1)] = \frac{\sigma_\varepsilon^2 \max(m, m')}{n} := H^2 \end{aligned}$$

since the definition of $\bar{\varphi}_j$ implies that $\int \bar{\varphi}_j^2(x) f(x) dx = 1$. Next

$$\sup_{t \in B_{m', m}^f(0,1)} \text{Var}(\eta_1 t(X_1)) \leq \mathbb{E}[\eta_1^2] \sup_{t \in B_{m', m}^f(0,1)} \mathbb{E}[t^2(X_1)] \leq \sigma_\varepsilon^2 := v$$

since $\mathbb{E}[t^2(X_1)] = \|t\|_f^2$. Lastly

$$\sup_{t \in B_{m', m}^f(0,1)} \sup_{(u, x)} (|u| \mathbf{1}_{|u| \leq k_n} |t(x)|) \leq k_n \sup_{t \in B_{m', m}^f(0,1)} \sup_x |t(x)|.$$

For $t = \sum_{j=0}^{m-1} a_j \varphi_j$, we have $\|t\|_f^2 = \vec{a}' \Psi_m \vec{a} = \|\sqrt{\Psi_m} \vec{a}\|_{2, m}^2$. Thus, for any m ,

$$\begin{aligned} \sup_{t \in B_m^f(0,1)} \sup_x |t(x)| &\leq c_\varphi \sqrt{m} \sup_{\|\sqrt{\Psi_m} \vec{a}\|_{2, m} = 1} \|\vec{a}\|_{2, m} \\ &\leq c_\varphi \sqrt{m} \sup_{\|\vec{u}\|_{2, m} = 1} \|\sqrt{\Psi_m^{-1}} \vec{u}\|_{2, m} = c_\varphi \sqrt{m} \sqrt{\|\Psi_m^{-1}\|_{\text{op}}}. \end{aligned}$$

Under condition (13), we have

$$\sqrt{m} \sqrt{\|\Psi_m^{-1}\|_{\text{op}}} \leq (m \|\Psi_m^{-1}\|_{\text{op}}^2)^{1/4} m^{1/4} \leq \left(\mathfrak{c} \frac{n}{\log(n)} \right)^{1/4} m^{1/4}.$$

We can take

$$(33) \quad M_1 := c_\varphi k_n \left(\mathfrak{c} \frac{n}{\log(n)} \right)^{1/4} (m \vee m')^{1/4}.$$

Consequently, Talagrand Inequality (see Theorem 8.3) implies, for $p(m, m') = 8 \frac{\sigma_\varepsilon^2 \max(m, m')}{n}$, and denoting by $m^* := \max(m, m')$,

$$\mathbb{E} \left(\sup_{t \in B_{m, m'}^f(0,1)} [\nu_{n,1}^2(t) - \frac{1}{2} p(m, m')] \right)_+ \leq \frac{C_1}{n} \left(e^{-C_2 m^*} + \frac{k_n^2 \sqrt{n} (m^*)^{1/2}}{n} e^{-C_3 \frac{n^{1/4} (m^*)^{1/4}}{k_n}} \right).$$

So, we choose $k_n = n^{1/4}$ and we get,

$$\mathbb{E} \left(\sup_{t \in B_{m',m}^f(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m, m') \right)_+ \leq \frac{C'_1}{n} \left(\exp(-C_2 m^*) + (m^*)^{1/2} \exp(-C_3 (m^*)^{1/4}) \right).$$

By summing up all terms over $m' \in \mathcal{M}_n$, we deduce

$$(34) \quad \mathbb{E} \left(\sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} [\nu_{n,1}]^2(t) - p(m, \hat{m}^{\text{th}}) \right)_+ \leq \sum_{m'} \mathbb{E} \left(\sup_{t \in B_{m',m}^f(0,1)} [\nu_{n,1}]^2(t) - p(m, m') \right)_+ \leq \frac{C}{n}.$$

Let us now study the second term in (32). Denote by $N_n \leq n$ the dimension of the largest space of the collection. Then we have

$$\begin{aligned} \mathbb{E} \left[\left(\sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} \nu_{n,2}^2(t) \right)_+ \right] &\leq \sum_{j=1}^{N_n} \mathbb{E} [\langle \xi, \bar{\varphi}_j \rangle_n^2] = \sum_{j=1}^{N_n} \text{Var} \left(\frac{1}{n} \sum_{i=1}^n \xi_i \bar{\varphi}_j(X_i) \right) \\ &= \frac{1}{n} \sum_{j=1}^{N_n} \mathbb{E} [\xi_1^2] \mathbb{E} [\bar{\varphi}_j^2(X_1)] \leq \frac{N_n}{n} \mathbb{E} [\varepsilon_1^2 \mathbf{1}_{|\varepsilon_1| > k_n}] \\ &\leq \frac{N_n}{n} \frac{\mathbb{E} [|\varepsilon_1|^{2+p}]}{k_n^p} \leq C \frac{\mathbb{E} [\varepsilon_1^6]}{n}, \end{aligned}$$

where the last line follows from the Markov inequality and the choices $k_n = n^{1/4}$ and $p = 4$. This bound together with (34) plugged in (32) gives the result of Lemma 7.1. \square

Proof of Lemma 7.2. As the collection of models is nested, we have

$$\mathbb{P}(\Omega_n^c) \leq \sum_{m \in \mathcal{M}_n} \mathbb{P} \left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2} \right).$$

Then

$$\mathbb{P} \left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2} \right) = \mathbb{P} \left(\sup_{t \in S_m, \|t\|_f=1} \left| \frac{1}{n} \sum_{i=1}^n [t^2(X_i) - \mathbb{E}t^2(X_i)] \right| > \frac{1}{2} \right).$$

Moreover we have

$$\begin{aligned} \sup_{t \in S_m, \|t\|_f=1} \left| \frac{1}{n} \sum_{i=1}^n [t^2(X_i) - \mathbb{E}t^2(X_i)] \right| &= \sup_{\vec{x} \in \mathbb{R}^m, \|\sqrt{\Psi_m} \vec{x}\|_{2,m}=1} \left| \vec{x}' \widehat{\Psi}_m \vec{x} - \vec{x}' \Psi_m \vec{x} \right| \\ &= \sup_{\vec{x} \in \mathbb{R}^m, \|\sqrt{\Psi_m} \vec{x}\|_{2,m}=1} \left| \vec{x}' (\widehat{\Psi}_m - \Psi_m) \vec{x} \right| \\ &= \sup_{\vec{u} \in \mathbb{R}^m, \|\vec{u}\|_{2,m}=1} \left| \vec{u}' \sqrt{\Psi_m}^{-1} (\widehat{\Psi}_m - \Psi_m) \sqrt{\Psi_m}^{-1} \vec{u} \right| \\ &= \|\sqrt{\Psi_m}^{-1} (\widehat{\Psi}_m - \Psi_m) \sqrt{\Psi_m}^{-1}\|_{\text{op}} \\ &\leq \|\sqrt{\Psi_m}^{-1}\|_{\text{op}} \|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} \|\sqrt{\Psi_m}^{-1}\|_{\text{op}} \\ &= \|\Psi_m^{-1}\|_{\text{op}} \|\widehat{\Psi}_m - \Psi_m\|_{\text{op}}. \end{aligned}$$

As a consequence,

$$(35) \quad \begin{aligned} \mathbb{P} \left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2} \right) &\leq \mathbb{P} \left(\|\Psi_m^{-1}\|_{\text{op}} \|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} > \frac{1}{2} \right) \\ &= \mathbb{P} \left(\|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} > \frac{1}{2\|\Psi_m^{-1}\|_{\text{op}}} \right). \end{aligned}$$

We apply Proposition 2.2 and we get

$$\mathbb{P} \left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2} \right) \leq 2m \exp \left(-\frac{1}{4c_\varphi^2} \frac{n}{m\|\Psi_m^{-1}\|_{\text{op}} \|f\|_\infty \|\Psi_m^{-1}\|_{\text{op}} + \frac{1}{6}} \right).$$

Now using $a + b \leq 2 \max(a, b)$ and the condition in the definition of \mathcal{M}_n in (13) imply

$$\mathbb{P} \left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2} \right) \leq 2\mathfrak{c}/n^5,$$

and summing up the terms over \mathcal{M}_n , with the bound on the cardinality implied by (13), gives the result of Lemma 7.2. \square

Proof of Theorem 2.1. We first write that

$$\begin{aligned} \hat{b}_{\hat{m}} - b &= (\hat{b}_{\hat{m}} - b)\mathbf{1}_{\hat{m} \in \mathcal{M}_n} + (\hat{b}_{\hat{m}} - b)\mathbf{1}_{\hat{m} \notin \mathcal{M}_n} \\ &= (\hat{b}_{\hat{m}^{\text{th}}} - b)\mathbf{1}_{\hat{m} \in \mathcal{M}_n} + (\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}})\mathbf{1}_{\hat{m} \notin \mathcal{M}_n} + (\hat{b}_{\hat{m}^{\text{th}}} - b)\mathbf{1}_{\hat{m} \notin \mathcal{M}_n} \\ &= (\hat{b}_{\hat{m}^{\text{th}}} - b) + (\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}})\mathbf{1}_{\hat{m} \notin \mathcal{M}_n} \\ &= (\hat{b}_{\hat{m}^{\text{th}}} - b) + (\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}})\mathbf{1}_{\hat{m} \notin \mathcal{M}_n, \hat{m}^{\text{th}} \notin \widehat{\mathcal{M}}_n}. \end{aligned}$$

Thus

$$(36) \quad \mathbb{E}(\|\hat{b}_{\hat{m}} - b\|_n^2) \leq 2\mathbb{E}(\|\hat{b}_{\hat{m}^{\text{th}}} - b\|_n^2) + 2\mathbb{E}(\|\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}}\|_n^2 \mathbf{1}_{\hat{m}^{\text{th}} \notin \widehat{\mathcal{M}}_n}).$$

Now $\hat{m}^{\text{th}} \in \mathcal{M}_n$ and $\hat{m}^{\text{th}} \notin \widehat{\mathcal{M}}_n$ means $\hat{m}^{\text{th}} \|\Psi_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 < \mathfrak{c}n/\log(n)$ and $\hat{m}^{\text{th}} \|\widehat{\Psi}_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 \geq 4\mathfrak{c}n/\log(n)$. This implies, as

$$\begin{aligned} 4\mathfrak{c} \frac{n}{\log(n)} &\leq \hat{m}^{\text{th}} \|\widehat{\Psi}_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 \leq 2\hat{m}^{\text{th}} \|\Psi_{\hat{m}^{\text{th}}}^{-1} - \widehat{\Psi}_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 + 2\hat{m}^{\text{th}} \|\Psi_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 \\ &\leq 2\hat{m}^{\text{th}} \|\Psi_{\hat{m}^{\text{th}}}^{-1} - \widehat{\Psi}_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 + 2\mathfrak{c} \frac{n}{\log(n)}, \end{aligned}$$

that $\hat{m}^{\text{th}} \|\widehat{\Psi}_{\hat{m}^{\text{th}}}^{-1} - \Psi_{\hat{m}^{\text{th}}}^{-1}\|_{\text{op}}^2 \geq \mathfrak{c}n/\log(n)$. Let us denote by

$$\Delta_m = \{m \|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}}^2 > \mathfrak{c}n/\log(n)\}.$$

Thus

$$(37) \quad \begin{aligned} \mathbb{E}(\|\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}}\|_n^2 \mathbf{1}_{\hat{m}^{\text{th}} \notin \widehat{\mathcal{M}}_n}) &\leq \mathbb{E}(\|\hat{b}_{\hat{m}} - \hat{b}_{\hat{m}^{\text{th}}}\|_n^2 \mathbf{1}_{\Delta_{\hat{m}^{\text{th}}}}) \\ &\leq 2 \left(\sqrt{\mathbb{E}[b^4(X_1)]} + \sqrt{\mathbb{E}[\varepsilon_1^4]} \right) \sqrt{\mathbb{P}(\Delta_{\hat{m}^{\text{th}}})}, \end{aligned}$$

by proceeding similarly to (31). Now we have

$$\mathbb{P}(\Delta_{\hat{m}^{\text{th}}}) \leq \sum_{m \in \mathcal{M}_n} \mathbb{P}(\Delta_m) \leq \sum_{m=1}^n \mathbb{P}(\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}}^2 > \|\Psi_m^{-1}\|_{\text{op}}^2).$$

Now, we write the decomposition

$$\begin{aligned} & \mathbb{P} \left[\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}} \right] \\ &= \mathbb{P} \left[\left\{ \|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}} \right\} \cap \left\{ \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} < \frac{1}{2} \right\} \right] \\ & \quad + \mathbb{P} \left[\left\{ \|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}} \right\} \cap \left\{ \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} \geq \frac{1}{2} \right\} \right] \\ & \leq \mathbb{P} \left[\left\{ \|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}} \right\} \cap \left\{ \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} < \frac{1}{2} \right\} \right] \\ (38) \quad & \quad + \mathbb{P} \left[\|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} \geq \frac{1}{2} \right]. \end{aligned}$$

To control the second term of the right hand side of (38), we write

$$(39) \quad \mathbb{P} \left[\|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} \geq \frac{1}{2} \right] \leq \mathbb{P} \left[\|\Psi_m^{-1}\|_{\text{op}} \|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} \geq \frac{1}{2} \right]$$

and we recognize (35) for which we already proved a bound.

Next to control the first term on the right hand side of (38), we apply Theorem 8.1 (with $\mathbf{A} = \Psi_m$ and $\mathbf{B} = \widehat{\Psi}_m - \Psi_m$), which yields

$$\begin{aligned} & \mathbb{P} \left[\left\{ \|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}} \right\} \cap \left\{ \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} < \frac{1}{2} \right\} \right] \\ & \leq \mathbb{P} \left[\left\{ \frac{\|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} \|\Psi_m^{-1}\|_{\text{op}}^2}{1 - \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}}} > \|\Psi_m^{-1}\|_{\text{op}} \right\} \cap \left\{ \|\Psi_m^{-1}(\widehat{\Psi}_m - \Psi_m)\|_{\text{op}} < \frac{1}{2} \right\} \right] \\ (40) \quad & \leq \mathbb{P} \left[\|\widehat{\Psi}_m - \Psi_m\|_{\text{op}} > \frac{1}{2} \|\Psi_m^{-1}\|_{\text{op}}^{-1} \right], \end{aligned}$$

which corresponds to (35) again and is thus bounded by a term of order $1/n^4$. So starting from (38) and gathering (39) and (40) gives

$$\mathbb{P}(\Delta_{\hat{m}}) \leq \sum_{m=1}^n \mathbb{P}(\Delta_m) \leq \frac{c}{n^2},$$

by applying Proposition 2.2 as previously. Plugging this in (37) and the result in (36), together with the result of Theorem 7.1, gives the result of Theorem 2.1. \square

7.1.5. *Proof of Corollary 2.1.* We have the following sequence of inequalities, for any $m \in \mathcal{M}_n$ and t any element of S_m ,

$$\begin{aligned} \|\hat{b}_{\hat{m}} - b_A\|_f^2 &= \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\ &\leq 2\|\hat{b}_{\hat{m}} - t\|_f^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\ &\leq 4\|\hat{b}_{\hat{m}} - t\|_n^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\ &\leq 8\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Omega_n} + 8\|t - b_A\|_n^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \end{aligned}$$

where Ω_n is defined by (28). Therefore, using the result of Theorem 2.1 and $\mathbb{E}(\|t - b_A\|_n^2) = \|t - b_A\|_f^2$, we get that for all $m \in \mathcal{M}_n$ and for any $t \in S_m$,

$$(41) \quad \mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_f^2) \leq C_1 \left(\|t - b_A\|_f^2 + \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C_2}{n} + \mathbb{E} \left(\|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \right),$$

so only the last term is to be studied. First, recall that Lemma 7.2 implies that $\mathbb{P}(\Omega_n^c) \leq c/n^4$. Next, write that $\|\hat{b}_{\hat{m}} - b_A\|_f^2 \leq 2(\|\hat{b}_{\hat{m}}\|_f^2 + \|b_A\|_f^2)$ and

$$\|\hat{b}_{\hat{m}}\|_f^2 = \int \left(\sum_{j=0}^{\hat{m}-1} \hat{a}_j \varphi_j(x) \right)^2 f(x) dx = (\vec{\hat{a}}^{(\hat{m})})' \Psi_{\hat{m}} \vec{\hat{a}}^{(\hat{m})} \leq \|\Psi_{\hat{m}}\|_{\text{op}} \|\vec{\hat{a}}^{(\hat{m})}\|_{2, \hat{m}}^2.$$

First, under $\|\sum_{j=0}^m \varphi_j^2\|_\infty \leq c_\varphi^2 m$, we get

$$\begin{aligned} \|\Psi_{\hat{m}}\|_{\text{op}} &= \sup_{\|\vec{x}\|_{2, \hat{m}}=1} \vec{x}' \Psi_{\hat{m}} \vec{x} = \sup_{\|\vec{x}\|_{2, \hat{m}}=1} \int \left(\sum_{j=0}^{\hat{m}-1} x_j \varphi_j(u) \right)^2 f(u) du \\ &\leq \sup_{\|\vec{x}\|_{2, \hat{m}}=1} \int \left(\sum_{j=0}^{\hat{m}-1} x_j^2 \sum_{j=0}^{\hat{m}-1} \varphi_j^2(u) \right) f(u) du \leq c_\varphi^2 \hat{m} \end{aligned}$$

Next, $\|\vec{\hat{a}}^{(\hat{m})}\|_{2, \hat{m}}^2 = (1/n^2) \|\widehat{\Psi}_{\hat{m}}^{-1} \widehat{\Phi}'_{\hat{m}} \vec{Y}\|_{2, \hat{m}}^2 \leq (1/n^2) \|\widehat{\Psi}_{\hat{m}}^{-1} \widehat{\Phi}'_{\hat{m}}\|_{\text{op}}^2 \|\vec{Y}\|_{2, n}^2$ and

$$\|\widehat{\Psi}_{\hat{m}}^{-1} \widehat{\Phi}'_{\hat{m}}\|_{\text{op}}^2 = \lambda_{\max} \left(\widehat{\Psi}_{\hat{m}}^{-1} \widehat{\Phi}'_{\hat{m}} \widehat{\Phi}_{\hat{m}} \widehat{\Psi}_{\hat{m}}^{-1} \right) = n \lambda_{\max}(\widehat{\Psi}_{\hat{m}}^{-1}) = n \|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\text{op}}$$

Therefore, for $\hat{m} \in \widehat{\mathcal{M}}_n$, we get

$$\|\hat{b}_{\hat{m}}\|_f^2 \leq c_\varphi^2 \frac{\hat{m} \|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\text{op}}}{n} \left(\sum_{i=1}^n Y_i^2 \right) \leq C \left(\sum_{i=1}^n Y_i^2 \right).$$

Then as $\mathbb{E}[(\sum_{i=1}^n Y_i^2)^2] \leq n^2 \mathbb{E}(Y_1^4)$, we get

$$\mathbb{E}(\|\hat{b}_{\hat{m}}\|_f^2 \mathbf{1}_{\Omega_n^c}) \leq \sqrt{\mathbb{E}(\|\hat{b}_{\hat{m}}\|_f^4) \mathbb{P}(\Omega_n^c)} \leq C \mathbb{E}^{1/2}(Y_1^4) n \mathbb{P}^{1/2}(\Omega_n^c) \leq c'/n.$$

On the other hand $\mathbb{E}(\|b_A\|_f^2 \mathbf{1}_{\Omega_n^c}) \leq \|b_A\|_f^2 \mathbb{P}(\Omega_n^c) \leq c''/n^4$. Thus $\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c}) \leq c_1/n$ and plugging this in (41) ends the proof of Corollary 2.1. \square

7.2. Proofs of Section 4. We need results on Laguerre functions with index $\delta > -1$. The Laguerre polynomial with index δ , $\delta > -1$, and degree k is given by

$$L_k^{(\delta)}(x) = \frac{1}{k!} e^x x^{-\delta} \frac{d^k}{dx^k} (x^{\delta+k} e^{-x}).$$

We consider the Laguerre functions with index δ , given by

$$(42) \quad \ell_k^{(\delta)}(x) = 2^{(\delta+1)/2} \left(\frac{k!}{\Gamma(k+\delta+1)} \right)^{1/2} L_k^{(\delta)}(2x) e^{-x} x^{\delta/2},$$

and $\ell_k^{(0)} = \ell_k$. The family $(\ell_k^{(\delta)})_{k \geq 0}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}^+)$.

In the following, we use the result of Askey and Wainger (1965) which gives bounds on ℓ_k , depending on k : for $\nu = 4k + 2\delta + 2$, and k large enough, it holds

$$|\ell_k^{(\delta)}(x/2)| \leq C \begin{cases} a) & (x\nu)^{\delta/2} & \text{if } 0 \leq x \leq 1/\nu \\ b) & (x\nu)^{-1/4} & \text{if } 1/\nu \leq x \leq \nu/2 \\ c) & \nu^{-1/4}(\nu-x)^{-1/4} & \text{if } \nu/2 \leq x \leq \nu - \nu^{1/3} \\ d) & \nu^{-1/3} & \text{if } \nu - \nu^{1/3} \leq x \leq \nu + \nu^{1/3} \\ e) & \nu^{-1/4}(x-\nu)^{-1/4} e^{-\gamma_1 \nu^{-1/2}(x-\nu)^{3/2}} & \text{if } \nu + \nu^{1/3} \leq x \leq 3\nu/2 \\ f) & e^{-\gamma_2 x} & \text{if } x \geq 3\nu/2 \end{cases}$$

where γ_1 and γ_2 are positive and fixed constants.

We need similar results for Hermite functions. These can be deduced from the following link between Hermite and Laguerre functions, proved in Comte and Genon-Catalot (2017):

Lemma 7.3. *For $x \geq 0$,*

$$h_{2n}(x) = (-1)^n \sqrt{x/2} \ell_n^{(-1/2)}(x^2/2), \quad h_{2n+1}(x) = (-1)^n \sqrt{x/2} \ell_n^{(1/2)}(x^2/2).$$

This is completed by the fact that Hermite functions are even for even n and odd for odd n .

7.2.1. Proof of Proposition 4.1. Under the assumption of the Proposition, for j large enough,

$$(43) \quad \int \varphi_j^2(x) f(x) dx \geq \frac{c_1}{\sqrt{j}},$$

where c_1 is a constant. The proof of Inequality (43) in the Hermite case is given in Belomestny *et al.* (2017), Proposition 2.2. and in Comte and Genon-Catalot (2017) in the Laguerre case.

Now we prove (20). As Ψ_m is a symmetric positive definite matrix, $\|\Psi_m^{-1}\|_{\text{op}} = 1/\lambda_{\min}(\Psi_m)$, where $\lambda_{\min}(\Psi_m)$ denotes the smallest eigenvalue of Ψ_m . By (14), we get that for all $j \in \{1, \dots, m\}$, denoting by \vec{e}_j the j th canonical vector (all coordinates are 0 except the j th which is equal to 1), $\vec{e}_j' \Psi_m \vec{e}_j = \int \varphi_j^2 f$, and

$$\min_{\|\vec{u}\|_{2,m}=1} \vec{u}' \Psi_m \vec{u} \leq \min_{j=1, \dots, m} \vec{e}_j' \Psi_m \vec{e}_j = \min_{j=1, \dots, m} \int \varphi_j^2 f \leq \frac{c}{\sqrt{m}}.$$

As a consequence, $\lambda_{\min}(\Psi_m) \leq c/\sqrt{m}$ which implies the result. \square

7.2.2. *Proof of Proposition 4.2.* We treat the Laguerre basis first. The result of Askey and Wainger (1965) recalled above states that, for k large enough, $\ell_k(x) \leq ce^{-\gamma_2 x}$ for $2x \geq 3(2k+1)$, where γ_2 is a constant. Thus for $\vec{x} \in \mathbb{R}^m$, $\|\vec{x}\|_{2,m} = 1$, we have

$$\begin{aligned} \vec{x}' \Psi_m \vec{x} &= \int_0^{+\infty} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 f(u) du \\ &\geq \int_0^{3(2m+1)} \left(\sum_{k=0}^{m-1} x_k \ell_k(v/2) \right)^2 f(v/2) dv/2 \\ &\geq \inf_{v \in [0, 3(2m+1)]} f(v/2) \int_0^{3(2m+1)/2} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 du \\ &\geq \inf_{u \in [0, 3(m+1/2)]} f(u) \left(\int_0^{+\infty} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 du - \int_{3(m+1/2)}^{+\infty} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 du \right) \end{aligned}$$

Then $\inf_{u \in [0, 3(m+1/2)]} f(u) \geq Cm^{-k}$ and $\int_0^{+\infty} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 du = \|\vec{x}\|_{2,m}^2 = 1$ and, for m large enough,

$$\int_{3(m+1/2)}^{+\infty} \left(\sum_{k=0}^{m-1} x_k \ell_k(u) \right)^2 du \leq C' m e^{-\gamma_3 m} \leq \frac{1}{2}.$$

It follows that, for m large enough, $\vec{x}' \Psi_m \vec{x} \geq Cm^{-k}/2$.

For the Hermite basis, we proceed analogously using that $|h_k(x)| \leq c|x|e^{-\gamma_2 x^2}$ for $x^2 \geq (3/2)(4k+3)$. \square

7.3. Proofs of Section 5.

7.3.1. *Proof of Proposition 5.1.* Consider the coupling method and the associated variables (X_i^*) with Berbee's Lemma, see Berbee (1979), with the method described in Vienet (1997, Prop.5.1 and its proof p.484).

Assume for simplicity that $n = 2p_n q_n$ for integers p_n, q_n . Then there exist random variables X_i^* , $i = 1, \dots, n$ satisfying the following properties:

- For $\ell = 0, \dots, p_n - 1$, the random vectors

$$\vec{X}_{\ell,1} = (X_{2\ell q_n+1}, \dots, X_{(2\ell+1)q_n})' \text{ and } \vec{X}_{\ell,1}^* = (X_{2\ell q_n+1}^*, \dots, X_{(2\ell+1)q_n}^*)'$$

have the same distribution, and so have the random vectors

$$\vec{X}_{\ell,2} = (X_{(2\ell+1)q_n+1}, \dots, X_{(2\ell+2)q_n})' \text{ and } \vec{X}_{\ell,2}^* = (X_{(2\ell+1)q_n+1}^*, \dots, X_{(2\ell+2)q_n}^*)'.$$

- For $\ell = 0, \dots, p_n - 1$,

$$(44) \quad \mathbb{P} \left[\vec{X}_{\ell,1} \neq \vec{X}_{\ell,1}^* \right] \leq \beta_{q_n} \text{ and } \mathbb{P} \left[\vec{X}_{\ell,2} \neq \vec{X}_{\ell,2}^* \right] \leq \beta_{q_n}.$$

- For each $\delta \in \{1, 2\}$, the random vectors $\vec{X}_{0,\delta}^*, \dots, \vec{X}_{p_n-1,\delta}^*$ are independent.

Then let $\Omega^* = \{X_i = X_i^*, i = 1, \dots, n\}$ and write that

$$\mathbb{P} \left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u \right] \leq \mathbb{P} \left[\{\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\} \cap \Omega^* \right] + \mathbb{P}[(\Omega^*)^c].$$

Then using the definition of the variables X_i^* , we get

$$\mathbb{P}[(\Omega^*)^c] \leq 2p_n \beta_{q_n} \leq c n e^{-\theta q_n}.$$

Then choosing $q_n = 5 \log(n)/\theta$ yields $\mathbb{P}[(\Omega^*)^c] \leq c/n^4$.

Now, we have to apply Tropp's result. To that aim, we write $\mathbf{S}_m = (1/2)(\mathbf{S}_{m,1} + \mathbf{S}_{m,2})$ where \mathbf{S}_m is given by (25), $\mathbf{S}_{m,1}$ is built with the $\vec{X}_{\ell,1}$:

$$\mathbf{S}_{m,1} = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \frac{1}{q_n} \sum_{r=1}^{q_n} \mathbf{K}_m(X_{2\ell q_n+r}) - \mathbb{E}(\mathbf{K}_m(X_{2\ell q_n+r}))$$

and $\mathbf{S}_{m,2}$ is analogously defined with with the $\vec{X}_{\ell,2}$. We have

$$\begin{aligned} \mathbb{P} \left[\{\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\} \cap \Omega^* \right] &= \mathbb{P} \left[\{\|\mathbf{S}_{m,1} + \mathbf{S}_{m,2}\|_{\text{op}} \geq 2u\} \cap \Omega^* \right] \\ &\leq \mathbb{P} \left[\{\|\mathbf{S}_{m,1}\|_{\text{op}} \geq u\} \cap \Omega^* \right] + \mathbb{P} \left[\{\|\mathbf{S}_{m,2}\|_{\text{op}} \geq u\} \cap \Omega^* \right] \\ &\leq \mathbb{P} \left[\|\mathbf{S}_{m,1}^*\|_{\text{op}} \geq u \right] + \mathbb{P} \left[\|\mathbf{S}_{m,2}^*\|_{\text{op}} \geq u \right], \end{aligned}$$

where $\mathbf{S}_{m,\delta}^*$, for $\delta = 1, 2$ are built on the $\vec{X}_{\ell,\delta}^*$. The two terms are similar, and we treat only the first one.

We can apply Tropp's result as $\mathbf{S}_{m,1}^*$ is a sum of p_n independent matrices. It follows from (26) that

$$\frac{1}{p_n q_n} \left\| \sum_{r=1}^{q_n} \mathbf{K}_m(X_{2\ell q_n+r}^*) - \mathbb{E}(\mathbf{K}_m(X_{2\ell q_n+r}^*)) \right\|_{\text{op}} \leq c_\varphi^2 \frac{m}{p_n} = \frac{10}{\theta} c_\varphi^2 \frac{m \log(n)}{n}.$$

Next, we must bound the variance of $\mathbf{S}_{m,1}^*$. We have

$$\nu(\mathbf{S}_{m,1}^*) = \frac{1}{p_n} \sup_{\|\vec{x}\|_{2,m}=1} \mathbb{E} \left(\frac{1}{q_n^2} \left\| \left[\sum_{r=1}^{q_n} (\mathbf{K}_m(X_r^*) - \mathbb{E}(\mathbf{K}_m(X_r^*))) \right] \vec{x} \right\|_{2,m}^2 \right)$$

Next,

$$\mathbb{E}_1 = \mathbb{E} \left(\frac{1}{q_n^2} \left\| \left[\sum_{r=1}^{q_n} (\mathbf{K}_m(X_r^*) - \mathbb{E}(\mathbf{K}_m(X_r^*))) \right] \vec{x} \right\|_{2,m}^2 \right) = \frac{1}{q_n^2} \sum_{j=0}^{m-1} \text{Var} \left[\sum_{r=1}^{q_n} \sum_{k=0}^{m-1} \varphi_j(X_r) \varphi_k(X_r) x_k \right]$$

Using the β -mixing variance inequality (see Viennet (1997), Theorem 2.1 p.472 and Lemma 4.2 p.481), there exists a nonnegative measurable function b_{q_n} such that for $p \geq 1$,

$$(45) \quad \mathbb{E}(b_{q_n}^p(X_1)) \leq \sum_{k \geq 0} (k+1)^{p-1} \beta_k,$$

and

$$\begin{aligned}\mathbb{E}_1 &\leq \frac{4}{q_n} \sum_{j=0}^{m-1} \int \left[\sum_{k=0}^{m-1} \varphi_j(u) \varphi_k(u) x_k \right]^2 b_{q_n}(u) f(u) du \\ &\leq \frac{4}{q_n} \sum_{j=0}^{m-1} \sum_{k=0}^{m-1} \int \varphi_j^2(u) \varphi_k^2(u) b_{q_n}(u) f(u) du\end{aligned}$$

Lemma 7.4. *If $\mathbb{E}(X^{8/3}) < +\infty$ for the Hermite basis and if $\mathbb{E}(1/X^2) < +\infty$ for the Laguerre basis, and $\sum_k (k+1)\beta_k < +\infty$, then*

$$\sum_{j=0}^{m-1} \sum_{k=0}^{m-1} \int \varphi_j^2(u) \varphi_k^2(u) b_{q_n}(u) f(u) du \leq Cm$$

where C depends on $\mathbb{E}(X^{8/3})$ or $\mathbb{E}(1/X^2)$ and $\sum_k k\beta_k$.

Then Lemma 7.4 implies that $\nu(\mathbf{S}_m) \leq Cm/n$ and applying Theorem 8.2 gives the announced result. \square

7.3.2. *Proof of Lemma 7.4.* We use the result of Askey and Wainger (1965) as recalled above. We set $\nu_j = 4j + 2$, $\nu_k = 4k + 2$.

We assume $j \leq k$, and for instance

$$0 \leq 1/\nu_k \leq 1/\nu_j \leq \nu_j/2 \leq \nu_k/2 \leq \nu_j - \nu_j^{1/3} \leq \nu_k - \nu_k^{1/3} \leq \nu_j + \nu_j^{1/3} \leq \nu_k + \nu_k^{1/3} \leq 3\nu_j/2 \leq 3\nu_k/2,$$

and we integrate $\ell_j^2 \ell_k^2$ on each subintervals.

- $\int_0^{1/\nu_k} \ell_j^2(u) \ell_k^2(u) b_{q_n}(u) f(u) du \lesssim \int_0^{1/\nu_k} b_{q_n}(u) f(u) du \lesssim \frac{1}{\nu_k} \int \frac{1}{u} b_{q_n}(u) f(u) du \leq \frac{C}{k+1}$
where $C = (\mathbb{E}(1/X^2) \sum_k (k+1)\beta_k)^{1/2}$, see (45). Now $\sum_{k=0}^{m-1} \sum_{j=0}^k (1/(k+1)) = m$.

- $\int_{1/\nu_k}^{1/\nu_j} \ell_j^2(u) \ell_k^2(u) b_{q_n}(u) f(u) du \lesssim \int_{1/\nu_k}^{1/\nu_j} \frac{1}{\sqrt{ku}} b_{q_n}(u) f(u) du \lesssim \frac{1}{\sqrt{\nu_k \nu_j}} \int \frac{1}{u} b_{q_n}(u) f(u) du \leq \frac{C}{\sqrt{(j+1)(k+1)}}$. Therefore $\sum_{k=0}^{m-1} \sum_{j=0}^k (1/\sqrt{(k+1)(j+1)}) \lesssim m$.

- $\int_{1/\nu_j}^{\nu_j/2} \ell_j^2(u) \ell_k^2(u) b_{q_n}(u) f(u) du \lesssim \frac{1}{\sqrt{\nu_j \nu_k}} \int_{1/\nu_j}^{\nu_j/2} \frac{1}{u} b_{q_n}(u) f(u) du \lesssim \frac{C}{\sqrt{(j+1)(k+1)}}$.

- $\int_{\nu_j/2}^{\nu_k/2} \ell_j^2(u) \ell_k^2(u) b_{q_n}(u) f(u) du \lesssim \frac{1}{\sqrt{\nu_j \nu_k}} \int_{\nu_j/2}^{\nu_k/2} \frac{1}{(\nu_j - u)^{1/4} \sqrt{u}} b_{q_n}(u) f(u) du$
 $\lesssim \frac{1}{\sqrt{(k+1)(j+1)^{1/2+1/6}}} \int_{\nu_j/2}^{\nu_k/2} \frac{1}{\sqrt{u}} b_{q_n}(u) f(u) du$.

Here $\sum_{k=0}^{m-1} \sum_{j=0}^k \frac{1}{\sqrt{(k+1)(j+1)^{1/2+1/6}}} \lesssim m^{5/6}$.

- $\int_{\nu_k/2}^{\nu_j - \nu_j^{1/3}} \ell_j^2(u) \ell_k^2(u) b_{q_n}(u) f(u) du \lesssim \frac{1}{\sqrt{\nu_j \nu_k} \nu_k^{1/6}} \int_{\nu_k/2}^{\nu_j - \nu_j^{1/3}} \frac{1}{\sqrt{u}} b_{q_n}(u) f(u) du$ and the sum

over j, k is also of order $m^{5/6}$.

The other terms are of lower order.

In the Hermite case, we proceed analogously using Lemma 7.3. \square

7.3.3. Proof of Theorem 5.1. We follow the line of the proof of Theorem 2.1, and we have to extend Lemma 7.1 and Lemma 7.2 to the dependent case.

For Lemma 7.2, the extension is the following. Note that we consider specifically Laguerre or Hermite bases but no longer require that $\|f\|_\infty < +\infty$.

Lemma 7.5. *Assume that $(X_i)_{i \geq 1}$ is strictly stationary geometrically β -mixing, with common density f and consider the Laguerre or Hermite basis. Then, for all $m \in \overline{\mathcal{M}}_n$ (see (23) and Ω_n defined by (28)), $\mathbb{P}(\Omega_n^c) \leq c/n^4$ where c is a positive constant.*

Proof of Lemma 7.5. We start from (35) and apply Proposition 5.1. We get

$$\mathbb{P}\left(\exists t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| > \frac{1}{2}\right) \leq 2m \exp\left(-\frac{1}{4a} \frac{n}{m \|\Psi_m^{-1}\|_{\text{op}}} \frac{1}{2\|\Psi_m^{-1}\|_{\text{op}} + 2\log(n)}\right) + \frac{c}{n^4}.$$

Using the definition of \mathcal{M}_n^* , we obtain the result. \square

Now we can extend Lemma 7.1 as follows.

Lemma 7.6. *Let $(X_i, i = 1, \dots, n)$ be observations from model (21) with $\mathbb{E}(\varepsilon_1^6) < +\infty$, let \hat{m}^{th} be defined by (27) using $\overline{\mathcal{M}}_n$ and consider the Laguerre or Hermite bases. Then $\bar{\nu}_n(t) = n^{-1} \sum_{i=1}^n \varepsilon_{i+1} t(X_i)$ satisfies*

$$\mathbb{E}\left(\sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} \bar{\nu}_n^2(t) - \bar{p}(m, \hat{m}^{\text{th}})\right)_+ \leq \frac{C}{n}$$

where $\bar{p}(m, m') = c\sigma_\varepsilon^2 \max(m, m')/n$ for c a numerical constant.

Proof of Lemma 7.6. We start with the same decomposition as in the proof of Lemma 7.1 and split $\bar{\nu}_n(t)$ into the sum $\bar{\nu}_{n,1} + \bar{\nu}_{n,2}$ as previously. The treatment of $\bar{\nu}_{n,2}$ is identical as it relies on a non-correlation property which is still true. We obtain the same bound with $k_n = (n/\log^2(n))^{1/4}$ and the maximal dimension $N_n \leq n/\log^2(n)$.

For $\bar{\nu}_{n,1}$ we proceed by the coupling strategy used in the proof of Proposition 5.1, applied to $u_i = (\varepsilon_{i+1}, X_i)$ which is also a β -mixing sequence with mixing coefficient such that $\beta_k \leq ce^{-\theta k}$, as in Baraud *et al.* (2001a). We denote by $\Omega^* = \{u_i = u_i^*, i = 1, \dots, n\}$. We still have $\mathbb{P}((\Omega^*)^c) \leq p_n \beta_{q_n} \leq c/n^4$ for $q_n = 5 \log(n)/\theta$.

On Ω^* , we replace the u_i by the u_i^* and split the term between odd and even blocks. We have to bound, say

$$\mathbb{E}\left(\sup_{t \in B_{\hat{m}^{\text{th}}, m}^f(0,1)} (\bar{\nu}_{n,1}^{*,1})^2(t) - \bar{p}(m, \hat{m}^{\text{th}})\right)_+$$

by using Talagrand inequality applied to mean of p_n independent random variables

$$\bar{\nu}_{n,1}^{*,1}(t) = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \left(\frac{1}{q_n} \sum_{r=1}^{q_n} \eta_{2\ell q_n+r}^* t(X_{2\ell q_n+r}^*) \right).$$

Clearly,

$$\mathbb{E}\left(\sup_{t \in B_{m,m'}^f(0,1)} (\bar{\nu}_{n,1}^{*,1})^2(t)\right) \leq \sigma_\varepsilon^2 \frac{\max(m, m')}{n} := H_\star^2$$

still holds. We have

$$\begin{aligned} \sup_{t \in B_{m,m'}^f(0,1)} \text{Var}\left(\frac{1}{q_n} \sum_{r=1}^{q_n} \eta_r^* t(X_r^*)\right) &= \sup_{t \in B_{m,m'}^f(0,1)} \text{Var}\left(\frac{1}{q_n} \sum_{r=1}^{q_n} \eta_r t(X_r)\right) \\ &= \frac{1}{q_n} \sup_{t \in B_{m,m'}^f(0,1)} \mathbb{E}(\eta_1^2) \mathbb{E}(t^2(X_1)) \leq \frac{\mathbb{E}(\varepsilon_1^2)}{q_n} := v_\star. \end{aligned}$$

Lastly

$$\sup_{t \in B_{m',m}^f(0,1)} \sup_{\vec{u}, \vec{x} \in \mathbb{R}^{q_n}} \left(\frac{1}{q_n} \sum_{r=1}^{q_n} |u_r| \mathbf{1}_{|u_r| \leq k_n} |t(x_r)| \right) \leq ck_n \left(\frac{n}{\log^2(n)} \right)^{1/4} (m \vee m')^{1/4} := M_1^\star,$$

where M_1^\star is computed analogously to M_1 given by (33), except that $m \in \overline{\mathcal{M}}_n$ increases the power of the log. Therefore, by applying Theorem 8.3, we obtain

$$\mathbb{E}\left(\sup_{t \in B_{m,m'}^f(0,1)} (\bar{\nu}_{n,1}^{*,1})^2(t) - 2H_\star^2\right)_+ \leq C_1 \left(\frac{1}{n} e^{-C_2(m \vee m')} + \frac{1}{n} (m \vee m')^{1/2} \exp(-C_3(m \vee m')^{1/4}) \right),$$

and

$$\mathbb{E}\left(\sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} (\bar{\nu}_{n,1}^{*,1})^2(t) - \bar{p}(m, \hat{m}^{\text{th}})\right)_+ \leq c/n.$$

It remains to bound

$$\mathbb{E}\left[\left(\sup_{t \in B_{\hat{m}^{\text{th}},m}^f(0,1)} (\bar{\nu}_{n,1})^2(t) - \bar{p}(m, \hat{m}^{\text{th}})\right) \mathbf{1}_{(\Omega^\star)^c}\right]_+.$$

We use the infinite norm computed to evaluate M_1^\star together with the bound on $\mathbb{P}[(\Omega^\star)^c]$ to obtain the result. \square

8. THEORETICAL TOOLS

A proof of the following theorem can be found in [Stewart and Sun, 1990].

Theorem 8.1. *Let \mathbf{A}, \mathbf{B} be $(m \times m)$ matrices. If \mathbf{A} is invertible and $\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}} < 1$, then $\tilde{\mathbf{A}} := \mathbf{A} + \mathbf{B}$ is invertible and it holds*

$$\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|_{\text{op}} \leq \frac{\|\mathbf{B}\|_{\text{op}} \|\mathbf{A}^{-1}\|_{\text{op}}^2}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}$$

Theorem 8.2 (Bernstein Matrix inequality). *Consider a finite sequence $\{\mathbf{S}_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that*

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

Introduce the random matrix $\mathbf{Z} = \sum_k \mathbf{S}_k$. Let $\nu(\mathbf{Z})$ be 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

$$\mathbb{E}\|\mathbf{Z}\|_{\text{op}} \leq \sqrt{2\nu(\mathbf{Z}) \log(d_1 + d_2)} + \frac{1}{3}L \log(d_1 + d_2).$$

Furthermore, for all $t \geq 0$

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

A proof can be found in [Tropp, 2012] or [Tropp, 2015].

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

Theorem 8.3. 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 , H and v such that $\sup_{f \in \mathcal{F}} \|f\|_{\infty} \leq M$, $\mathbb{E}[\sup_{f \in \mathcal{F}} |\nu_n(f)|] \leq H$, and $\sup_{f \in \mathcal{F}} (1/n) \sum_{i=1}^n \text{Var}(f(X_i)) \leq v$.

Then for all $\alpha > 0$,

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

with $C(\alpha) = (\sqrt{1 + \alpha} - 1) \wedge 1$, 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 $f \rightarrow \nu_n(f)$ is continuous and \mathcal{F} contains a countable dense family.

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.
- [Aksey and Wainger, 1965] Aksey, R. and Wainger, S. (1965) Mean convergence of expansions in Laguerre and Hermite series. *Amer. J. Math.* **87**, 695-708.
- [Baraud, 2000] Baraud, Y. (2000) Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117**, 467-493.
- [Baraud, 2002] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.
- [Baraud et al., 2001a] Baraud, Y., Comte, F. and Viennet, G. (2001a) Adaptive estimation in autoregression or β -mixing regression via model selection. *Ann. Statist.* **29**, 839-875.
- [Baraud et al., 2001b] Baraud, Y., Comte, F. and Viennet, G. (2001b) Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.* **5**, 33-49.
- [Belomestny et al., 2016] Belomestny, D., Comte, F., and Genon-Catalot, V. (2016). Nonparametric Laguerre estimation in the multiplicative censoring model. *Electron. J. Statist.*, 10(2):3114–3152.
- [Belomestny et al., 2017] Belomestny, D., Comte, F., and Genon-Catalot, V. (2017). Sobolev-Hermite versus Sobolev nonparametric density estimation on R To appear in *The Annals of the Institute of Mathematical Statist.*
- [Berbee, 1979] Berbee, H.C.P. (1979). Random walks with stationary increments and renewal theory. *Math. Tracts. Mathematisch Centrum*, Amsterdam, **112**.
- [Bongioanni and Torrea, 2006] Bongioanni, B. and Torrea, J.L. (2006). Sobolev spaces associated to the harmonic oscillator. *Proc.Indian Acad. Sci. (math. Sci.)* **116** (3), 337-360.

- [Bongioanni and Torrea, 2009] Bongioanni, B. and Torrea, J. L. (2009). What is a Sobolev space for the Laguerre function systems? *Studia Math.*, 192(2):147–172.
- [Brunel and Comte, 2009] Brunel, E. and Comte, F. (2009) Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.* **3**, 1-24.
- [Chen et al., 2012] Chen, R. Y., Gittens, A., and Tropp, J. A. (2012). The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Inf. Inference*, 1(1):2–20.
- [Comte et al., 2015] Comte, F., Cuenod, C.-A., Pensky, M., and Rozenholc, Y. (2017). Laplace deconvolution and its application to dynamic contrast enhanced imaging. *J. R. Stat. Soc., Ser. B*, **79**, 69-94.
- [Comte and Genon-Catalot, 2015] Comte, F. and Genon-Catalot, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models. *Electron. J. Stat.*, **9**, 1112-1148.
- [Comte and Genon-Catalot, 2017] Comte, F. and Genon-Catalot, V. (2017). Laguerre and Hermite bases for inverse problems. *Preprint MAP5 2017-05*.
- [Comte et al., 2007] Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007) Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514-543.
- [deVore and Lorentz, 1993] DeVore, R.A. and Lorentz, G.G. (1993) *Constructive approximation*, Springer-Verlag, Berlin.
- [Efromovich, 1999] Efromovich, S. (1999) *Nonparametric curve estimation. Methods, theory, and applications*. Springer Series in Statistics. Springer-Verlag, New York.
- [Mabon, 2017] Mabon, G. (2017). Adaptive deconvolution on the nonnegative real line. *Scandinavian Journal of Statistics*, 44:707-740.
- [Plancade, 2011] Plancade, S. (2011) Model selection for hazard rate estimation in presence of censoring. *Metrika* **74**, 313-347.
- [Shen, 2000] Shen, J. (2000). Stable and efficient spectral methods in unbounded domains using Laguerre functions. *SIAM J. Numer. Anal.* **38**, 1113-1133.
- [Stewart and Sun, 1990] Stewart, G. W. and Sun, J.-G. (1990). *Matrix perturbation theory*. Boston etc.: Academic Press, Inc.
- [Szegő, 1975] Szegő, 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.
- [Vareschi, 2015] Vareschi, T. (2015). Noisy Laplace deconvolution with error in the operator. *J. Stat. Plann. Inference*, 157–158:16–35.
- [Viennet, 1997] Viennet, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467-492.
- [Walter, 1977] Walter, G.G. (1977). Properties of Hermite series estimation of probability density. *Annals of Statistics*, **5**, 1258-1264.

APPENDIX A. NUMERICAL ILLUSTRATIONS

In this section, numerical illustrations of how our method works are presented. The estimation procedure is implemented for the Laguerre (Figures 2 to 5) and the Hermite basis (Figure 6). The $(\varepsilon_i)_{1 \leq i \leq n}$ are generated as an i.i.d. sample of Gaussian $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$. Then, we choose different functions $b(\cdot)$ (bounded or not) and different types of distribution of the design $(X_i)_{1 \leq i \leq n}$. Typically, a linear function $x \mapsto 2x + 1$ is experimented without the information of its linearity, which allows to test moment conditions; on the contrary, $x \mapsto 4x/(1+x^2)$ is bounded and should be easier to reconstruct. For the design density, we consider standard uniform or Gaussian cases, and also different heavy tailed distributions.

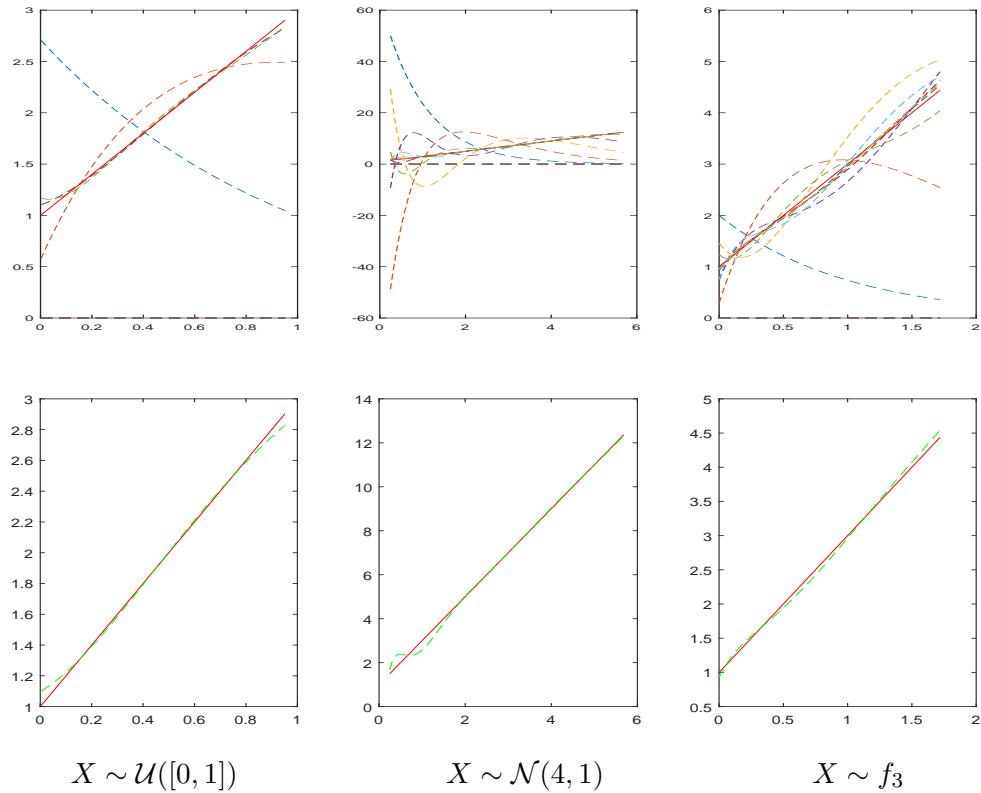


FIGURE 2. First line: beam of the proposals \hat{f}_m for $m = 1$ to m_{\max} in the Laguerre basis. Second line: the estimator as selected by the procedure, $\hat{f}_{\hat{m}}$. Function $b(x) = 2x + 1$, $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$.

In Figure 2, we plot in the first line the collection of estimators in the Laguerre basis, among which the algorithm makes the selection. The number of computed estimators is different from one example to another, as the collection of models $\widehat{\mathcal{M}}_n$ is random and depends on $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$. In the practical implementation, we consider the (random) maximum value m_{\max} such that $\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \leq n$, since inversion of the matrix $\widehat{\Psi}_m$ remains possible in such cases. Surprisingly, we can see that very few estimators are sometimes computed

(see the example of uniform distribution on the right). They are also very different from one dimension to another. The second line presents the final estimator, selected by the procedure. In the example of Figure 1, the curve is linear, and is perfectly estimated, although its particular form is unknown and was not *a priori* easy to obtain with the Laguerre basis.

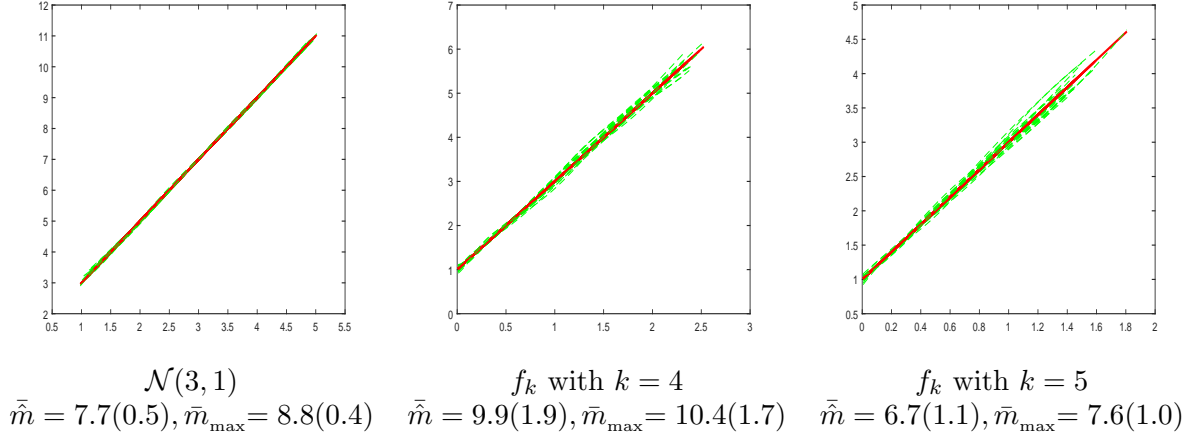


FIGURE 3. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x + 1$ and different laws for the design, $f_k(x) = (k - 1)/(1 + x)^k \mathbf{1}_{x \geq 0}$.

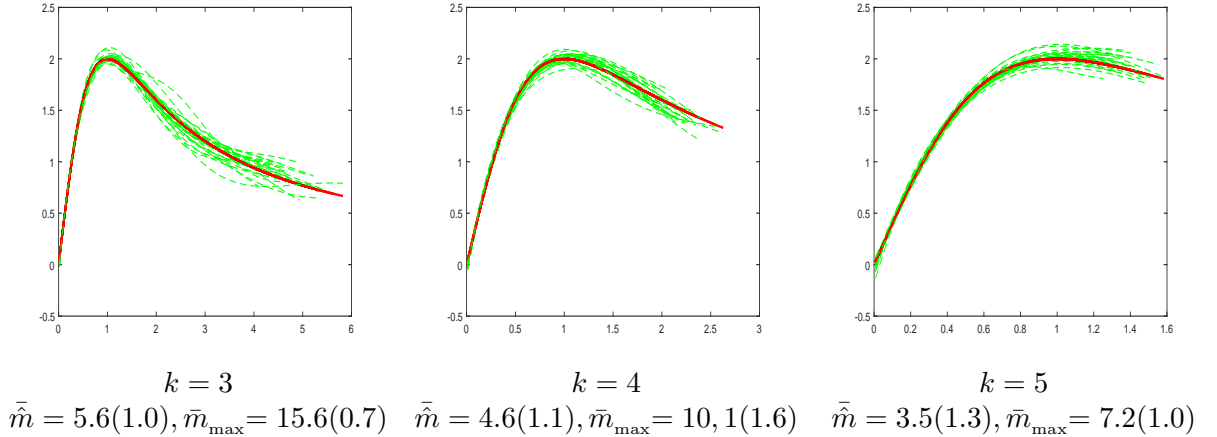


FIGURE 4. 25 estimated curves in the Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, density $f_k(x) = (k - 1)/(1 + x)^k \mathbf{1}_{x \geq 0}$ for $k = 3, 4$ and 5 , $b(x) = 4x/(1 + x^2) \mathbf{1}_{x \geq 0}$.

In Figures 3, 4 and 5, we present beams of 25 estimators computed in the Laguerre basis, they give information about the variability of the procedure. Figure 3 is complementary of Figure 2 and considers the same linear regression function with similar distributions for

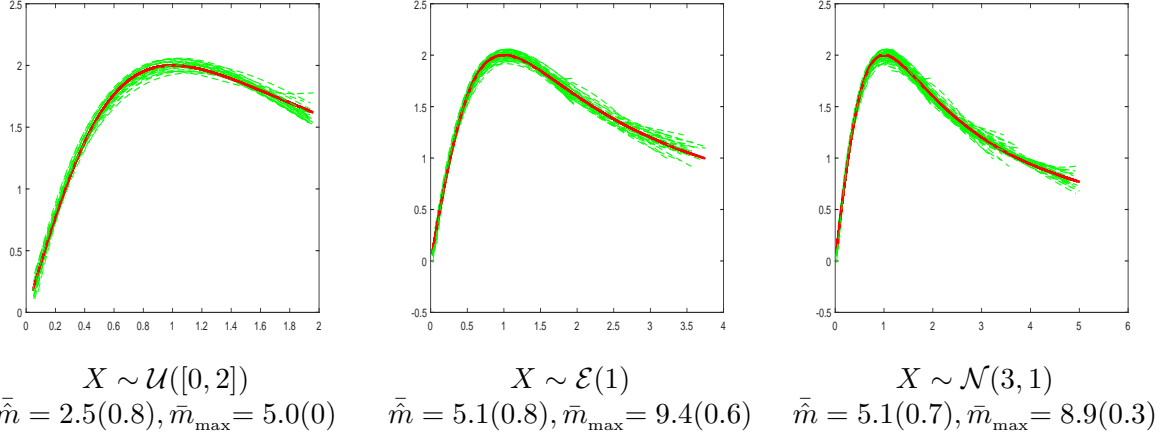


FIGURE 5. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 4x/(1+x^2)\mathbf{1}_{x \geq 0}$ and different laws for the design.

X , and Figure 4 presents the results for the function $b(x) = 4x/(1+x^2)\mathbf{1}_{x \geq 0}$ and different heavy tailed distributions for X . The beams illustrate the stability of the algorithm, with some design distributions leading to better results, probably due to higher signal-to-noise ratio. The interest of the linear case is also to illustrate the sharpness of the moment conditions: indeed the condition $\mathbb{E}[b^2(X_1)] < +\infty$ for X with density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$ is satisfied for $k > 3$ and the condition $\mathbb{E}[b^4(X_1)] < +\infty$ holds for $k > 5$. We checked that the method does not work for $k = 2, 3$, but the last two plots of Figure 3 show that it works rather well for $k = 4, 5$. The minimal theoretical condition may thus be weakened from $\mathbb{E}[b^4(X_1)] < +\infty$ to $\mathbb{E}[b^2(X_1)] < +\infty$. The Hermite basis has similar behaviour and an example is provided in Figure 6.

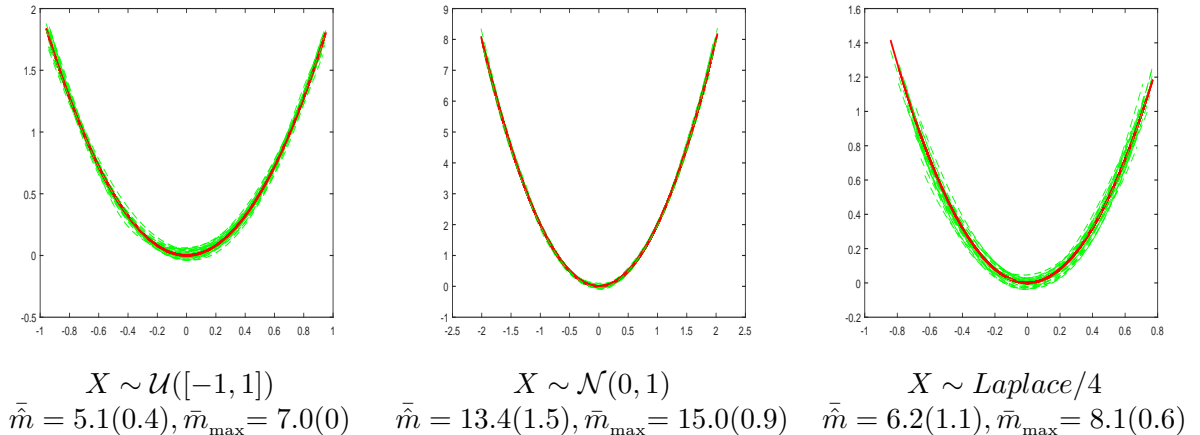


FIGURE 6. 25 estimated curves in Hermite basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x^2$ and different laws for the design.

Below each plot, we give the density of the design and the value of $\bar{\hat{m}}$ which is the mean of the selected dimensions for the 25 estimators represented on the figure, with standard deviation in parenthesis. It is associated with the value of \bar{m}_{\max} which is the mean of the maximal dimension for which the estimator is computed, with standard deviation in parenthesis. We can see that the maximal dimension is rather small (less than ten models are compared for selection, in general) but an adequate choice seems always to exist in this small collection. This means that the squared-bias variance compromise in the restricted set \mathcal{M}_n has good performance and that the non compact Laguerre and Hermite bases are very interesting and simple estimation tools. Indeed, the method is very fast and this low complexity, already argued in Belomestny *et al.* (2017), has an important practical interest.