**Annals of the Institute of Statistical Mathematics manuscript No.**
(will be inserted by the editor)

# Regression function estimation as a partly inverse problem

**F. Comte · V. Genon-Catalot**

Received: date / Revised: date

**Abstract** This paper is about nonparametric regression function estimation. 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 most previous results. Upper and lower bounds for resulting rates are provided.

**Keywords** Hermite basis · Laguerre basis · Model selection · Non parametric estimation · Regression function

## 1 Introduction

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

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

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

**The online version of this article contains supplementary material.**

F. Comte, Corresponding author
MAP5 UMR CNRS 8145, University Paris Descartes,
45, rue des Saints-Pères, 75006 Paris, France
E-mail: fabienne.comte@parisdescartes.fr

V. Genon-Catalot
MAP5 UMR CNRS 8145, University Paris Descartes,
45, rue des Saints-Pères, 75006 Paris, France
E-mail: valentine.genon-catalot@parisdescartes.fr

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, two smoothing parameters.

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

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

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. It provides directly an estimator of $b$ restricted to the set $A$, a unique 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, 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 $f$. These arguments are in favour of the second strategy. Noting that the least squares contrast can be rewritten

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

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. Moreover, the equivalence of the norms $\|t\|_n$ and $\|t\|_f$ for $t$ in a finite dimensional linear space must be handled. This is done by Cohen *et al.* (2013) and we take advantage of their findings.

However, Cohen *et al.* (2013)'s work has drawbacks: their stability condition is settled in terms of an unknown quantity; the regression function is assumed to be bounded by a known quantity and the definition of the estimator depends on this known bound; they do not study the model selection problem. Due to their statistically simplified setting, they do not deal with the entire partially inverse problem hidden in the procedure.

Our aim in this work is to obtain theoretical results in regression function estimation by the least squares projection method described above, and we want to handle the case of possibly non compact support $A$ of the basis. This explains why we must avoid boundedness assumption on $b$. A consequence is that the cutoff which has to be introduced in the definition of the estimator depends on the behaviour of the eigenvalues of a random matrix. This requires a specific study to obtain a bound on the integrated $\mathbb{L}^2$ risk, and makes the model selection question near of an inverse problem with unknown operator.

What is the interest of non compactly supported bases? In general, the estimation set and the bases support are considered as fixed in the theoretical part, while are in practice adjusted on the data. With a non compact support, it is not necessary to fix a preliminary definition. Moreover, we have at disposal non compactly supported bases such as the Laguerre ($A = \mathbb{R}^+$) or the Hermite ($A = \mathbb{R}$) basis which have been used recently for nonparametric estimation by projection (see *e.g.* Comte and Genon-Catalot, 2015, 2018, Belomestny *et al.* 2016), showing that theses bases are both convenient and with specific properties. They are especially useful in certain inverse problems (see Comte *et al.* 2017, Mabon, 2017).

Before giving our plan, let us highlight our main findings.
- First, we propose a new procedure of estimation relying on a random cutoff, and generalize Cohen *et al.* (2013)'s results, with a more statistical flavour.
- We deduce from the bias-variance decomposition upper rates of the estimator on specific Sobolev spaces, for which lower bounds are also established. We recover the standard rates of the "compact case" but also exhibit non standard ones when considering Laguerre or Hermite bases and spaces.
- We propose a model selection procedure for regression function estimation on a set $A$ whether compact or not, where the collection of models itself is random and prove that it reaches automatically a bias-variance tradeoff. We highlight the regression problem as a partially inverse problem: the eigenvalues of the matrix which must be inverted play a role in the problem not directly as a weight on the variance term but in the definition of the collection of models.

The framework and plan of the paper is the following. We fix a set $A \subset \mathbb{R}$ and concentrate on the estimation of the regression function $b$ restricted to a set $A$, $b_A := b\mathbf{1}_A$. As $A$ may be unbounded, we do not want to assume that $b_A \in \mathbb{L}^2(A, dx)$ which would exclude linear or polynomial functions. Our main assumption is that $b_A \in \mathbb{L}^4(A, f(x)dx)$, i.e. $\mathbb{E}b_A^4(X_1) < +\infty$ which is rather weak. In Section 2, we define the projection estimator of the regression function $b_A$ and check that the most elementary risk bound holds without any constraint on the support $A$ or the projection basis. In Section 3, we prove a risk-bound for the estimator on one model, borrowing some elements to Cohen *et al.* (2013)'s results to extend them. Then, we study rates and optimality for the integrated $\mathbb{L}^2(A, f(x)dx)$-risk. Introducing regularity spaces linked with $f$, we prove upper and matching lower bounds for our projection estimator. Then we quickly show how to recover existing results for compactly supported bases and more precisely illustrate the case of non compact support with the Hermite and Laguerre bases for estimation on $A = \mathbb{R}$ and $A = \mathbb{R}^+$ respectively. In Section 4, 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 both for the integrated empirical risk and for the integrated $\mathbb{L}^2(A, f(x)dx)$-risk: it generalizes existing results to non compactly supported bases. Section 5 gives some concluding remarks. Most proofs are gathered in Section 6. An appendix in Supplementary material gives theoretical tools used along the proofs and presents numerical illustrations of the method.

## 2 Projection estimator and preliminary results

Recall that $f$ denotes the density of $X_1$. In the following, $\|.\|_{2,p}$ denotes the euclidean norm in $\mathbb{R}^p$. For $A \subset \mathbb{R}$, $\|.\|_A$ denotes the integral norm in $\mathbb{L}^2(A, dx)$, $\|.\|_f$ the integral norm in $\mathbb{L}^2(A, f(x)dx)$ and $\|.\|_\infty$ the supremum norm on $A$. For any function $h$, $h_A = h\mathbf{1}_A$.

### 2.1 Definition of the projection estimator

Consider model (1). Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 0, \ldots, m-1)$ be an orthonormal system of $A$-supported functions belonging to $\mathbb{L}^2(A, dx)$. Define $S_m = \text{span}(\varphi_0, \ldots, \varphi_{m-1})$, the linear space spanned by $(\varphi_0, \ldots, \varphi_{m-1})$. Note that the $\varphi_j$'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). 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 $\langle \mathbf{u}, t\rangle_n = \frac{1}{n}\sum_{i=1}^n u_i t(X_i)$ when $\mathbf{u}$ is the vector $(u_1, \ldots, u_n)'$, $\mathbf{u}'$ denotes the transpose of $\mathbf{u}$ and $t$ is a function. We introduce the matrices

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

and

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

Set $\mathbf{Y} = (Y_1, \dots, Y_n)'$ and define $\hat{\mathbf{a}}^{(m)} = (\hat{a}_0^{(m)}, \dots, \hat{a}_{m-1}^{(m)})'$ as the $m$-dimensional vector such that $\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j^{(m)} \varphi_j$. Assuming that $\widehat{\Psi}_m$ is invertible almost surely (a.s.) yields:

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

### 2.2 Bound on the mean empirical risk on a fixed space

We now evaluate the risk of the estimator, without any constraint on the basis support. Though classical, the result hereafter requires noteworthy comments.

**Proposition 1** *Let $(X_i, Y_i)_{1 \le i \le n}$ 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 $\widehat{\Psi}_m$ is a.s. invertible. Consider the least squares estimator $\hat{b}_m$ of $b$, given by (4). Then*

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

$$\le \inf_{t \in S_m} \left[\int (b_A - t)^2(x) f(x) dx\right] + \sigma_\varepsilon^2 \frac{m}{n}. \tag{6}$$

It is not obvious from (6) or from the previous formula that the bias term is small when $m$ is large. Therefore, two questions arise: is $\Psi_m$ invertible for any $m$, and does the bias tend to zero when $m$ grows to infinity? The Lemmas below provide sufficient conditions. These conditions can be refined if the basis is specified.

**Lemma 1** *Assume that $\lambda(A \cap \mathrm{supp}(f)) > 0$ where $\lambda$ is the Lebesgue measure and $\mathrm{supp}(f)$ the support of $f$, that the $(\varphi_j)_{0 \le j \le m-1}$ are continuous, and that there exist $x_0, \dots, x_{m-1} \in A \cap \mathrm{supp}(f)$ such that $\det[(\varphi_j(x_k))_{0 \le j,k \le m-1}] \neq 0$. Then, $\Psi_m$ is invertible.*

**Lemma 2** *Assume that $b_A \in \mathbb{L}^2(A, f(x)dx)$. Assume that $(\varphi_j)_{j \ge 0}$ is an orthonormal basis of $\mathbb{L}^2(A, dx)$ such that, for all $j \ge 0$, $\int \varphi_j^2(x) f(x) dx < +\infty$, that $f$ is bounded on $A$ and that for all $x \in A$, $f(x) > 0$.*
*Then $\inf_{t \in S_m} \left[\int (b_A - t)^2(x) f(x) dx\right]$ tends to 0 when $m$ tends to infinity.*

Lemma 1 follows from the following equality. For all $\mathbf{u} = (u_0, \dots, u_{m-1})' \in \mathbb{R}^m \setminus \{\mathbf{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j \varphi_j(x)$, $\mathbf{u}' \Psi_m \mathbf{u} = \|t\|_f^2 = \int_A t^2(x) f(x) dx \ge 0$. Under the assumptions, the result follows.
The proof of Lemma 2 is elementary. 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, the system $\phi_j = \varphi_j \sqrt{f}$, $j \ge 0$ is a complete family of $\mathbb{L}^2(A, dx)$. Indeed, if $g \in \mathbb{L}^2(A, dx)$, $\int g\phi_j = \int \varphi_j (g\sqrt{f}) = 0 \; \forall j \ge 0$ implies $g = 0$ using our assumptions.
The result of Proposition 1 is general in the sense that it holds for any basis support, whether compact or not. We stress that (5) is an equality, in particular the variance term is **exactly** equal to $\sigma_\varepsilon^2 m/n$. In addition, the result does not depend on the basis.

**Remark 1** *We underline that the latter fact is not obvious. Consider the density estimation setting, where $\hat{f}_m = \sum_{j=0}^{m-1} \widehat{c}_j \varphi_j$ with $\widehat{c}_j = (1/n)\sum_{i=1}^n \varphi_j(X_i)$ is a projection estimator of $f$. Then the integrated $\mathbb{L}^2-$risk bound is*

$$\mathbb{E}(\|\hat{f}_m - f_A\|^2) = \inf_{t\in S_m} \|f_A - t\|^2 + \frac{\sum_{j=0}^{m-1}\mathbb{E}[\varphi_j^2(X_1)]}{n} - \frac{\|f_m\|^2}{n},$$

*where $f_m = \sum_{j=0}^{m-1}\langle f, \varphi_j\rangle\varphi_j$ is the $\mathbb{L}^2(dx)$-orthogonal projection of $f$ on $S_m$. The variance term has the order of $\sum_{j=0}^{m-1}\mathbb{E}[\varphi_j^2(X_1)]/n$. For most compactly supported bases, this term has order $m/n$ (for instance, it is equal to $m/n$ for histograms or trigonometric polynomial basis, see section 3.3); but it is proved in Comte and Genon-Catalot (2018) that for Laguerre or Hermite basis (see section 3.4 below), this term has exactly the order $\sqrt{m}/n$ (lower and upper bound are provided, under weak assumptions). This is why it is important to see that, in regression context, the variance order does not depend on the basis.*

### 2.3 Useful inequalities

For $M$ a matrix, we denote by $\|M\|_{\rm op}$ the operator norm defined as the square root of the largest eigenvalue of $MM'$. If $M$ is symmetric, it coincides with $\sup\{|\lambda_i|\}$ where $\lambda_i$ are the eigenvalues of $M$. Moreover, if $M, N$ are two matrices with compatible product $MN$, then, $\|MN\|_{\rm op} \leq \|M\|_{\rm op}\|N\|_{\rm op}$.
The possible values of the dimension $m$ to study the collection $(\hat{b}_m)$ of estimators are subject to restrictions, for which the following property is important:

**Proposition 2** *Assume that the spaces $S_m$ are nested (i.e. $m \leq m' \Rightarrow S_m \subset S_{m'}$) and $\Psi_m$ (resp. $\widehat{\Psi}_m$) is invertible, then $m \mapsto \|\Psi_m^{-1}\|_{\rm op}$ (resp $m \mapsto \|\widehat{\Psi}_m^{-1}\|_{\rm op}$) is nondecreasing.*

Let us define

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

This quantity is independent of the choice of the $\mathbb{L}^2(dx)$-orthonormal basis of $S_m$, and for nested spaces $S_m$, the map $m \mapsto L(m)$ is increasing. We need to study the set

$$\Omega_m(\delta) = \left\{ \sup_{t\in S_m,\, t\neq 0} \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \delta \right\} \tag{8}$$

where the empirical and the $\mathbb{L}^2(A, f(x)dx)$ norms are equivalent on $S_m$. Theorem 1 in Cohen *et al.* (2013) provides the adequate inequality. In our context, it takes the following form:

**Proposition 3** *Let $\widehat{\Psi}_m$, $\Psi_m$ be the $m \times m$ matrices defined in Equation* (3) *and assume that $\Psi_m$ is invertible. Then for all $0 \le \delta \le 1$,*

$$\mathbb{P}(\Omega_m(\delta)^c) = \mathbb{P}\left[\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \delta\right] \le 2m\exp\left(-c(\delta)\frac{n}{L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1)}\right),$$

*where $\mathrm{Id}_m$ denotes the $m \times m$ identity matrix and $c(\delta) = \delta + (1-\delta)\log(1-\delta)$.*

By convention, we set $\|\Psi_m^{-1}\|_{\mathrm{op}} = +\infty$ whenever $\Psi_m$ is not invertible. As a consequence, if $m$ is such that

$$L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1) \le \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}, \quad \mathfrak{c} = \frac{1-\log(2)}{5}, \tag{9}$$

we obtain that, choosing $\delta = 1/2$, the set $\Omega_m := \Omega_m(1/2)$ satisfies $\mathbb{P}(\Omega_m^c) \le 2n^{-4}$. Indeed, $L(m)\|\Psi_m^{-1}\|_{\mathrm{op}} \ge m$ (see Lemma 4 in the proof of Proposition 3). Condition (9) can be understood as ensuring the stability of the least-squares estimator, as underlined in Cohen *et al.* (2013). We can prove:

**Proposition 4** (i) *Assume that $f$ is bounded. Let $\widehat{\Psi}_m$ be the $m \times m$ matrix defined in Equation* (3)*. Then for all $u > 0$*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \ge u\right] \le 2m\exp\left(-\frac{n\,u^2/2}{L(m)\ (\|f\|_\infty + 2\,u/3)}\right).$$

(ii) *Assume that $\widehat{\Psi}_m, \Psi_m$ are (a.s.) invertible. Then for $\alpha > 0$,*

$$\left\{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \alpha\|\Psi_m^{-1}\|_{\mathrm{op}}\right\} \subset \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \frac{\alpha \wedge 1}{2}\right\}.$$

## 3 Truncated estimator on a fixed space

We may consider from Proposition 1 that the problem is standard. However, it is known that difficulties arise if we want to bound the integrated $\mathbb{L}^2$-risk instead of the empirical risk, even for fixed $m$. Actually, the general regression problem is an inverse problem since the link between the function of interest $b$ and the density of the observations $(Y_i, X_i)_i$ is of convolution type $f_Y(y) = \int f_\varepsilon(y - b(x))f(x)dx$ where $f_Y$ and $f_\varepsilon$ are the densities of $Y_1$ and $\varepsilon_1$. This can also be seen from the fact that the estimator is computed via the inversion of the matrix $\widehat{\Psi}_m$. Thus we can expect that the procedure depends on the eigenvalues of $\Psi_m$.

## 3.1 Integrated risk bound

Let us assume as above that $b_A \in \mathbb{L}^2(A, f(x)dx)$. It is not possible to deduce from Proposition 1 a bound on $\mathbb{E}[\|\hat{b}_m - b_A\|_f^2]$ for all $m$ such that $\widehat{\Psi}_m$ is invertible. On the other hand, we introduce a cutoff and define

$$\tilde{b}_m := \hat{b}_m \mathbf{1}_{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \vee 1) \leq \mathfrak{c}n/\log(n)}, \tag{10}$$

where $L(m)$ is defined by (7) and $\mathfrak{c}$ in (9). On the set $\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \vee 1) \leq \mathfrak{c}n/\log(n)\}$, the matrix $\widehat{\Psi}_m$ is invertible and its eigenvalues $(\lambda_i)_{1\leq i\leq m}$ satisfy $\inf_{1\leq i\leq m}(\lambda_i) \geq m\log(n)/(\mathfrak{c}n)$. Analogously, condition (9) is equivalent to the fact that $\Psi_m$ is invertible and its eigenvalues are lower bounded by $2m\log(n)/(\mathfrak{c}n)$. We have:

**Proposition 5** *Assume that $\mathbb{E}(\varepsilon_1^4) < +\infty$ and $b_A \in \mathbb{L}^4(A, f(x)dx)$. Then for any $m$ satisfying (9), we have*

$$\mathbb{E}[\|\tilde{b}_m - b_A\|_f^2] \leq \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right) \inf_{t\in S_m} \|b_A - t\|_f^2 + 8\sigma_\varepsilon^2 \frac{m}{n} + \frac{c}{n}, \tag{11}$$

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

The proof of Proposition 5 exploits as a first step the proof of Theorem 3 in Cohen *et al.* (2013). However, the estimator in Cohen *et al.* (2013) is mainly theoretical: indeed they assume that $b$ is bounded and the estimator depends on the bound, which has thus to be known. As $A$ may be unbounded, it is important to get rid of this restriction.

## 3.2 Rate and optimality

So far, the bias rate of the $\mathbb{L}^2(A, f(x)dx)$-risk in (6) and (11) has not been assessed. To this end, we introduce regularity spaces related to $f$ by setting:

$$W_f^s(A, R) = \left\{h \in \mathbb{L}^2(A, f(x)dx), \forall \ell \geq 1, \|h - h_\ell^f\|_f^2 \leq R\ell^{-s}\right\} \tag{12}$$

where we recall that $h_\ell^f$ is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of $h$ on $S_\ell$. From (11), we easily deduce an upper bound for the risk, which we state below. The risk rate is optimal, as we also prove the following lower bound.

**Theorem 1** *Assume that $b_A \in W_f^s(A, R)$ and that $m_{\text{opt}} := [n^{1/(s+1)}]$ satisfies (9).*

- *Upper bound. If $\mathbb{E}(\varepsilon_1^4) < +\infty$, $\mathbb{E}(\|\tilde{b}_{m_{\text{opt}}} - b_A\|_f^2) \leq Cn^{-s/(s+1)}$.*
- *Lower bound. If $\varepsilon_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2)$,*

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

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

The condition that $m_{\rm opt} = [n^{1/(s+1)}]$ satisfies (9) is actually mainly a constraint on $f$, see the discussion at the end of Section 3.4.
The partly inverse problem appears here. The rate of $\|\Psi_m^{-1}\|_{\rm op}$ as a function of $m$ is to be interpreted as a measure of the degree of ill-posedness of the inverse problem, in the context of regression function estimation.

**Proposition 6** *Under the assumptions of Theorem 1, if moreover $L(m) \lesssim m$ and $\|\Psi_m^{-1}\|_{\rm op} \lesssim m^k$, then $\mathbb{E}[\|\hat{b}_{m_{\rm opt}} - b_A\|_f^2] \leq C(R)n^{-\frac{s}{(s\vee k)+1}}$.*

This result is due to the fact that the constraint $L(m)\|\Psi_m^{-1}\|_{\rm op} \lesssim m^{k+1} \lesssim n/\log(n)$ has to be fulfilled for $m_{\rm opt}$.

### 3.3 Case of compact $A$ and compactly supported bases

In this section, we assume that $A$ is compact and give examples of bases where, for simplicity, $A = [0,1]$. 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 $j = 1,\dots,(m-1)/2$.

For spaces generated by histograms and by trigonometric basis, $L(m) = m$, for spaces generated by piecewise polynomials with degree $r$, $L(m) = (r+1)m$. Spaces generated by compactly supported wavelets also satisfy (7) with $L(m)$ of order $m$. The trigonometric spaces are nested; for histograms, piecewise polynomials and wavelets, the models are nested if the subdivisions are diadic ($m = 2^k$ for increasing values of $k$).

Let $P_k(x) = \sqrt{2}L_k(2x-1)\,\mathbf{1}_{[0,1]}(x)$, for $k = 0,\dots,m-1$ be the Legendre polynomial basis rescaled from $[-1,1]$ to $[0,1]$. It is an $\mathbb{L}^2([0,1],dx)$ orthonormal basis of $S_m = \text{span}(P_0,\dots,P_{m-1})$. As $\|P_k\|_\infty = \sqrt{2}\sqrt{2k-1}$, we get $L(m) = 2m^2$ (see Cohen *et al.* (2013)).
If $A$ is compact, one can assume that

$$\exists f_0 > 0, \text{ such that } \forall x \in A, \quad f(x) > f_0. \tag{13}$$

This assumption is commonly and crucially used in papers on nonparametric regression. In particular, it implies that $\Psi_m$ is invertible, and more precisely:

**Proposition 7** *Assume that Assumption (13) is satisfied, then*

$$\forall m \leq n, \quad \|\Psi_m^{-1}\|_{\rm op} \leq 1/f_0.$$

Indeed (13) implies that, for $\mathbf{u} = (u_0,\dots,u_{m-1})'$ a vector of $\mathbb{R}^m$, $\mathbf{u}'\Psi_m\mathbf{u}$ is equal to

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

Therefore $\|\Psi_m^{-1}\|_{\rm op} \leq 1/f_0$ and Proposition 7 is proved. A consequence of (13) is that the matrix $\Psi_m$ needs not appear in condition (9), thus the matrix $\widehat{\Psi}_m$ needs not appear in the definition of $\tilde{b}_m$. So we can define, as in Baraud (2002), for $c'$ a constant,

$$\tilde{b}_m = \hat{b}_m \mathbf{1}_{L(m)\leq c'n/\log(n)}. \tag{15}$$

Now, let us discuss about the usual rates in this compact setting. Assume that

$$b_A \in \mathbb{L}^2(A, dx) \text{ and } \|f\|_\infty < +\infty. \tag{16}$$

Then $\forall t \in S_m$, $\|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - t\|_A^2$ and thus

$$\inf_{t\in S_m} \|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - b_m\|_A^2 \tag{17}$$

where $b_m$ is the $\mathbb{L}^2(A, dx)$-orthogonal projection of $b_A$ on $S_m$. Thus we recover a classical bias, and the bias-variance compromise leads to standard rates, typically $n^{-2\alpha/(2\alpha+1)}$ for $b_A \in \mathcal{B}_{\alpha,2,\infty}(A, R)$ a Besov ball with radius $R$ and regularity $\alpha$ (see De Vore and Lorentz (1993), or Baraud (2002, section 2)).

### 3.4 Examples of non compact $A$ and non compactly supported bases

If $A$ is not compact, assumption (13) can not hold, therefore we can not get rid of the matrix $\Psi_m$. Our contribution is to take into account and enlight the role of $\Psi_m$ and to introduce a new selection procedure involving a random collection of models (see Section 4).

Now we assume that

$$b_A \in \mathbb{L}^2(A, f(x)dx), \quad \lambda(A \cap \text{supp}(f)) > 0, \text{ and } f \text{ is upper bounded.} \tag{18}$$

We give 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 (2018) for density estimation by projection using these bases.

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

$$L_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x\geq 0}, \quad j \geq 0. \tag{19}$$

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)):

$$\forall j \geq 0, \ \forall x \in \mathbb{R}^+, \ |\ell_j(x)| \leq \sqrt{2}. \tag{20}$$

Clearly, the collection of models $(S_m = \text{span}\{\ell_0, \ldots, \ell_{m-1}\})$ is nested, and (20) implies that this space satisfies (7) with $L(m) = 2m$ (the supremum is attained at $x = 0$).

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

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

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. The infinite norm of $h_j$ satisfies (see Abramowitz and Stegun (1964), Szegö (1975) p.242):

$$\|h_j\|_\infty \leq \Phi_0, \qquad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160, \tag{22}$$

so that the Hermite space satisfies (7) with $L(m) \leq \Phi_0^2 m$. The collection of models is also clearly nested.

Hereafter, we use the notation $\varphi_j$ to denote $\ell_j$ in the Laguerre case and $h_j$ in the Hermite case. We denote by $S_m = \text{span}(\varphi_0, \varphi_1, \ldots, \varphi_{m-1})$ the linear space generated by the $m$ functions $\varphi_0, \ldots, \varphi_{m-1}$ 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.
The matrices $\Psi_m$, $\widehat{\Psi}_m$ in these bases have specific properties:

**Lemma 3** *For all $m \in \mathbb{N}$, for all $m \leq n$, $\widehat{\Psi}_m$ is a.s. invertible.*

The result below on $\Psi_m$ is crucial for understanding our procedure.

**Proposition 8** *For all $m$, $\Psi_m$ is invertible and there exists a constant $c^\star$ such that,*

$$\|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^\star m. \tag{23}$$

In the Laguerre and Hermite cases, Inequality (23) 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 (9) leads to restrictions on the values $m$, as illustrated by the next proposition.

**Proposition 9** *Consider the Laguerre or the Hermite basis. Assume that $f(x) \geq c/(1+x)^k$ for $x \geq 0$ and $k \geq 2$ in the Laguerre case or $f(x) \geq c/(1+x^2)^k$ for $x \in \mathbb{R}$ and $k \geq 1$ in the Hermite case. Then for $m$ large enough, $\|\Psi_m^{-1}\|_{\text{op}} \leq Cm^k$.*

We performed numerical experiments which seem to indicate that the order $m^k$ is sharp.
If $f$ is as in Proposition 9 and $b_A \in W_f^s(A, R)$, then Proposition 6 applies: the optimal rate of order $n^{-s/(s+1)}$ can be reached by the adaptive estimator if $s > k$. Note that in a Sobolev-Laguerre ball:

$$W^s(\mathbb{R}^+, R) = \{h \in \mathbb{L}^2(A, dx), \quad \sum_{j \geq 0} j^s \langle h, \ell_j \rangle^2 \leq R\}, \tag{24}$$

the index $s$ (and not $2s$) 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)). The same type of property holds for Sobolev-Hermite balls, see Belomestny *et al.* (2019). Therefore, the rate $n^{-s/(s+1)}$ is non standard[1].

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 $n^{-2s/(2s+1)}$ (see Remark 1). It seems that, in the regression setting, we cannot have such a gain. Analogous considerations hold with the Hermite basis.

We do not know the order of $\|\Psi_m^{-1}\|_{\rm op}$ for $f$ exponential or Gaussian: it is likely to increase exponentially fast. However, the bias term is then also likely to decrease exponentially fast. Thus, the resulting risk may remain quite small: this is what we observe in simulations.

## 4 Adaptive procedure

Let us consider now the following assumptions.

**(A1)** The collection of spaces $S_m$ is nested (that is $S_m \subset S_{m'}$ for $m \leq m'$) and such that, for each $m$, the basis $(\varphi_0, \dots, \varphi_{m-1})$ of $S_m$ satisfies

$$\forall m \geq 1, \quad L(m) = \| \sum_{j=0}^{m-1} \varphi_j^2 \|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{25}$$

**(A2)** $\|f\|_\infty < +\infty$.

We present now a model selection procedure and associated risk bounds. To select the most relevant space $S_m$, we proceed by choosing

$$\hat{m} = \arg\min_{m \in \widehat{\mathcal{M}}_n} \left\{ -\|\hat{b}_m\|_n^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right\} \tag{26}$$

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

$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathbb{N}, m(\|\widehat{\Psi}_m^{-1}\|_{\rm op}^2 \vee 1) \leq \mathfrak{d} \frac{n}{\log(n)} \right\}, \quad \mathfrak{d} = \frac{1}{192\, c_\varphi^2 (\|f\|_\infty \vee 1 + (1/3))}. \tag{27}$$

The value of the constant $\mathfrak{d}$ is determined below by Lemma 7.
A theoretical counterpart of $\widehat{\mathcal{M}}_n$, with $\mathfrak{d}$ is defined in (27), is useful:

$$\mathcal{M}_n = \left\{ m \in \mathbb{N}, m\ (\|\Psi_m^{-1}\|_{\rm op}^2 \vee 1) \leq \frac{\mathfrak{d}}{4} \frac{n}{\log(n)} \right\}. \tag{28}$$

[1] If $b_A$ is a combination of $\Gamma$-type functions, then the bias term $\inf_{t \in S_m} \|b_A - t\|^2$ is much smaller (exponentially decreasing) and the rate $\log(n)/n$ can be reached by the adaptive estimator (see e.g. Mabon (2017)).

Note that the cutoff for defining $\hat{m}$ and $\hat{b}_{\hat{m}}$ is different from the one used in (10). As $m(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq m(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}^2 \vee 1)$, this yields a smaller set of possible values for $\hat{m}$.
The procedure (26) aims at performing an automatic bias-variance tradeoff. Each term is related to the bias or the variance obtained in Proposition 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 $b_A$ 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.

**Theorem 2** *Let $(X_i, Y_i)_{1\leq i\leq n}$ be observations from model (1). Assume that* **(A1)**, **(A2)** *hold, that $\mathbb{E}(\varepsilon_1^6) < +\infty$ and $\mathbb{E}\big[b^4(X_1)\big] < +\infty$. Then, there exists a numerical constant $\kappa_0$ such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2\big] \leq C \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \kappa\sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C'}{n}, \tag{29}$$

*and*

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_f^2\big] \leq C_1 \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \kappa\sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C_1'}{n} \tag{30}$$

*where $C, C_1$ are a numerical constants and $C', C_1'$ are constants depending on $\|f\|_\infty$, $\mathbb{E}[b^4(X_1)]$, $\mathbb{E}(\varepsilon_1^6)$.*

Theorem 2 shows that the risk of the estimator $\hat{b}_{\hat{m}}$ automatically realizes the bias-variance tradeoff, up to the multiplicative constants $C, C_1$, both in term of empirical and of integrated $\mathbb{L}^2(A, f(x)dx)-$risk. The conditions are general, rather weak, and do not impose any support constraint. Theorem 2 contains existing results when the bases are regular and compactly supported.

The numerical constant $\kappa$ provided by the theory (here $\kappa \geq \kappa_0 = 32$) is not optimal from theoretical point of view and too large in practice. It is thus standard to choose its value from preliminary calibration of the method, on simulated experiments. Here we took $\kappa = 4$, see the supplementary material.

The constant $\mathfrak{d}$ in the definition of $\widehat{\mathcal{M}}_n$ depends on $\|f\|_\infty$ which is unknown. In practice, this quantity may be replaced by an estimator, possibly rough. Otherwise, to avoid looking for the value of $\mathfrak{d}$, we can replace the bound $\mathfrak{d}n/\log(n)$ in $\widehat{\mathcal{M}}_n$ by $n/\log^{1+\epsilon}(n)$, $\epsilon > 0$, for $n$ is large enough.

The constant $\sigma_\varepsilon^2$ is also generally unknown, and must be replaced by an estimator. We simply propose to use the residual least-squares estimator: $\widehat{\sigma_\varepsilon^2} = (1/n)\sum_{i=1}^n (Y_i - \hat{b}_{m^*}(X_i))^2$ 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.

## 5 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 $\widehat{b} = \widehat{bf}/\widehat{f}$. 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 mostly 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.

Several extensions of our method can be obtained.

First, note that the result of Proposition 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 $\varepsilon_i$.

We also may have considered the heteroskedastic regression the model $Y_i = b(X_i) + \sigma(X_i)\varepsilon_i$, $\mathrm{Var}(\varepsilon_1) = \mathbb{E}(\varepsilon_1^2) = 1$, and the same contrast. The estimator on $S_m$ is still given by (4). Assuming that $\sigma^2(x)$ is uniformly bounded, we can obtain results similar to those obtained here.

Note that 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, see Bouaziz *et al.* (2018). 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 convenient that the Laguerre basis does not require to choose it.

## 6 Proofs

6.1 Proof of Proposition 1.

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

$$\|\hat{b}_m - b_A\|_n^2 = \|\Pi_m b - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2 = \inf_{t \in S_m} \|t - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2. \tag{31}$$

By taking expectation, we obtain

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] \le \inf_{t \in S_m} \int (t - b_A)^2(x) f(x) dx + \mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right]. \tag{32}$$

Now we have:

$$\mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2 \frac{m}{n}. \tag{33}$$

The result of Proposition 1 follows. □

**Proof of equality (33)**. Denote by $b(X) = (b(X_1), \dots, b(X_n))'$ and $b_A(X) = (b_A(X_1), \dots, b_A(X_n))'$. We can write

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

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

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

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

$$\|\hat{b}_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X)\boldsymbol{\varepsilon}\|_n^2 = \frac{1}{n} \boldsymbol{\varepsilon}' \mathbf{P}(X)' \mathbf{P}(X) \boldsymbol{\varepsilon} = \frac{1}{n} \boldsymbol{\varepsilon}' \mathbf{P}(X) \boldsymbol{\varepsilon} \tag{34}$$

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 $\mathbb{E}(\|\mathbf{P}(X)\boldsymbol{\varepsilon}\|_{2,n}^2) \le \mathbb{E}(\|\boldsymbol{\varepsilon}\|_{2,n}^2) < +\infty$. Next, we compute, using that $\mathbf{P}(X)$ has coefficients depending on the $X_i$'s only,

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

where Tr(.) is the trace of the matrix. So, we find

$$\mathrm{Tr}(\mathbf{P}(X)) = \mathrm{Tr}\big((\widehat{\Phi}'_m \widehat{\Phi}_m)^{-1} \widehat{\Phi}'_m \widehat{\Phi}_m\big) = \mathrm{Tr}(\mathrm{I}_m) = m$$

where $\mathrm{I}_m$ is the $m \times m$ identity matrix. Finally, we get $\mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2 (m/n)$. This is (33). □

### 6.2 Proof of Proposition 2

Let $t = \sum_{j=0}^{m-1} a_j \varphi_j$, and $\mathbf{a} = (a_0, \dots, a_{m-1})'$, then $\|t\|^2 = \|\mathbf{a}\|_{2,m} = \mathbf{a}'\,\mathbf{a}$ and $\|t\|_f^2 = \mathbf{a}'\Psi_m\mathbf{a} = \|\Psi_m^{1/2}\mathbf{a}\|_{2,m}^2$, where $\Psi_m^{1/2}$ is a symmetric square root of $\Psi_m$. Thus

$$\sup_{t\in S_m, \|t\|_f=1} \|t\|^2 = \sup_{\mathbf{a}\in\mathbb{R}^m, \|\Psi_m^{1/2}\mathbf{a}\|_{2,m}=1} \mathbf{a}'\,\mathbf{a}.$$

Set $\mathbf{b} = \Psi_m^{1/2}\mathbf{a}$, that is $\mathbf{a} = \Psi_m^{-1/2}\mathbf{b}$. Then

$$\sup_{t\in S_m, \|t\|_f=1} \|t\|^2 = \sup_{\mathbf{b}\in\mathbb{R}^m, \|\mathbf{b}\|_{2,m}=1} \mathbf{b}'\Psi_m^{-1}\mathbf{b} = \|\Psi_m^{-1}\|_{\mathrm{op}}.$$

As, for $m \le m'$, we assume $S_m \subset S_{m'}$, we also have

$$\|\Psi_m^{-1}\|_{\mathrm{op}} = \sup_{t\in S_m, \|t\|_f=1} \|t\|^2 \le \sup_{t\in S_{m'}, \|t\|_f=1} \|t\|^2 = \|\Psi_{m'}^{-1}\|_{\mathrm{op}}.$$

The same holds for $\sup_{t\in S_m, \|t\|_n=1} \|t\|^2 = \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$. □

### 6.3 Proof of Proposition 3

The first equality holds by writing

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

Now, Theorem 1 in Cohen *et al.* (2013) yields that for $0 < \delta < 1$, $\mathbb{P}(\Omega_m(\delta)^c) \le 2me^{-c(\delta)n/K(m)}$ where, for $(\theta_j)_{0\le j\le m-1}$ an $\mathbb{L}^2(A, f(x)dx)$-orthonormal basis of $S_m$,

$$K(m) = \sup_{x\in A} \sum_{j=0}^{m-1} \theta_j^2(x), \tag{35}$$

provided that $K(m) < +\infty$.[2] Note that $K(m)$ is independent of the choice of the basis $(\theta_j)_{0\le j\le m-1}$. Then, the following lemma:

**Lemma 4** *Assume that $\Psi_m$ is invertible and $L(m) < +\infty$ (see (7)). Then $K(m) < +\infty$, and for $\overrightarrow{\varphi_{(m)}}(x) = (\varphi_0(x), \dots, \varphi_{m-1}(x))'$, we have*

$$m \le K(m) = \sup_{x\in A} \overrightarrow{\varphi_{(m)}}(x)'\Psi_m^{-1}\overrightarrow{\varphi_{(m)}}(x) \le L(m)\|\Psi_m^{-1}\|_{\mathrm{op}}.$$

[2] In Cohen *et al.* (2013), the condition $K(m) < +\infty$ is not clearly stated; it is implicit as the result does not hold otherwise. Actually all examples of the paper are for $A$ compact, in which case $K(m) < +\infty$. If $A$ is not compact, then $K(m)$ may be $+\infty$. Therefore our condition (7) and Lemma 4 clarify Cohen *et al.*'s result.

gives the result of Proposition 3. □
**Proof of Lemma 4.** We have $\sum_{j=0}^{m-1}\int\theta_j^2(x)f(x)dx = m \leq K(m)$. Now, let $\overrightarrow{\theta_{(m)}}(x) = (\theta_0(x),\dots,\theta_{m-1}(x))'$. There exists an $m\times m$ matrix $A_m$ such that $\overrightarrow{\theta_{(m)}}(x) = A_m\overrightarrow{\varphi_{(m)}}(x)$. By definition of the basis $(\theta_j)_{0\leq j\leq m}$,

$$\int_A \overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)'f(x)dx = \mathrm{Id}_m$$

and

$$\int_A \overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)'f(x)dx = A_m\Psi_m A_m'.$$

This implies $A_m^{-1}(A_m')^{-1} = (A_m'A_m)^{-1} = \Psi_m$ and $A_m'A_m = \Psi_m^{-1}$. Thus

$$\overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)' = \overrightarrow{\varphi_{(m)}}(x)A_m'A_m\overrightarrow{\varphi_{(m)}}(x) = \overrightarrow{\varphi_{(m)}}(x)'\Psi_m^{-1}\overrightarrow{\varphi_{(m)}}(x).$$

This gives the first equality. To end the proof of Lemma 4, note that the last term is bounded by $\|\Psi_m^{-1}\|_{\mathrm{op}}\|\overrightarrow{\varphi_{(m)}}(x)\|_{2,m}^2 = \|\Psi_m^{-1}\|_{\mathrm{op}}\sum_{j=0}^{m-1}\varphi_j^2(x)$ . □
Note that we can see also here that $\mathbf{G}$ in Cohen *et al.* (2013), that we denote here $\widehat{\mathbf{G}}_m$ is such that $\widehat{\mathbf{G}}_m = A_m\widehat{\Psi}_m A_m'$ where $A_m'$ is a square root of $\Psi_m^{-1}$.

## 6.4 Proof of Proposition 4

**Proof of** `(i)`. To get the announced result, we apply again a Bernstein matrix inequality given in Tropp (2012)(see Theorem A.2 in Supplementary material). We write $\widehat{\Psi}_m$ as a sum of a sequence of independent matrices $\widehat{\Psi}_m = \dfrac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i)$, with $\mathbf{K}_m(X_i) = (\varphi_j(X_i)\varphi_k(X_i))_{0\leq j,k\leq m-1}$. We define

$$\mathbf{S}_m = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right]. \qquad (36)$$

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

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

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

$$\begin{aligned}\|\mathbf{K}_m(X_1)\|_{\mathrm{op}} &= \sup_{\|\mathbf{x}\|_{2,m}=1}\sum_{0\leq j,k\leq m-1} x_jx_k[\mathbf{K}_m(X_1)]_{j,k}\\ &= \sup_{\|\mathbf{x}\|_{2,m}=1}\sum_{0\leq j,k\leq m-1} x_jx_k\varphi_j(X_1)\varphi_k(X_1)\\ &= \sup_{\|\mathbf{x}\|_{2,m}=1}\left[\left(\sum_{j=0}^{m-1}x_j\varphi_j(X_1)\right)^2\right] \leq L(m).\end{aligned}$$

So we get that, a.s.

$$\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\rm op}/n \leq \frac{2L(m)}{n} := \mathbf{L}. \tag{37}$$

• Bound on $\nu(\mathbf{S}_m) = \| \sum_{i=1}^n \mathbb{E}\left[(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])' \left(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right]\right)\right] \|_{\rm op}/n^2$. We have

$$\nu(\mathbf{S}_m) = \frac{1}{n} \sup_{\|\mathbf{x}\|_{2,m}=1} \mathbb{E}\,\|\left(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\right)\mathbf{x}\|_{2,m}^2 .$$

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

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

Therefore as, by **(A2)**, $f$ is bounded,

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

Then we get that $\nu(\mathbf{S}_m) \leq \dfrac{\|f\|_\infty L(m)}{n}$. Applying Theorem A.2 in Supplementary material (see Tropp (2012)) gives the result `(i)` of Proposition 4. □

**Proof of `(ii)`.** First note that

$$\begin{aligned}\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\rm op} &= \|\Psi_m^{-1/2}\left(\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\right)\Psi_m^{-1/2}\|_{\rm op}) \\ &\leq \|\Psi_m^{-1}\|_{\rm op}\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\|_{\rm op},\end{aligned}$$

so that

$$\left\{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\rm op} > \alpha\|\Psi_m^{-1}\|_{\rm op}\right\} \subset \left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\|_{\rm op} > \alpha\right\}. \tag{38}$$

Now, we write the decomposition $\left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\|_{\rm op} > \alpha\right\} := B_1 \cup B_2$ with

$$\begin{aligned}B_1 &= \left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\|_{\rm op} > \alpha\right\} \bigcap \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\rm op} < \frac{1}{2}\right\} \\ B_2 &= \left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} - \mathrm{Id}_m\|_{\rm op} > \alpha\right\} \bigcap \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\rm op} \geq \frac{1}{2}\right\}.\end{aligned}$$

Clearly $B_2 \subset \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\rm op} \geq \dfrac{1}{2}\right\}$.

Applying Theorem A.1 (see Stewart and Sun (1990) and Theorem A.1 in supplementary material) with $\mathbf{A} = \mathrm{Id}_m$ and $\mathbf{B} = \Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m$, yields

$$
\begin{aligned}
B_1 &\subset \left\{ \frac{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}}{1 - \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}} > \alpha \right\} \cap \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} < \frac{1}{2} \right\} \\
&\subset \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \alpha/2 \right\} \cap \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} < \frac{1}{2} \right\} \\
&\subset \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \alpha/2 \right\}.
\end{aligned}
$$

Thus $B_1 \cup B_2 \subset \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq \frac{\alpha \wedge 1}{2} \right\}$, which ends the proof of `(ii)` and of Proposition 4. □

## 6.5 Proof of Proposition 5

We define the sets (see (10)),

$$
\Lambda_m = \left\{ L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c}\frac{n}{\log(n)} \right\}, \text{ and } \Omega_m = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \frac{1}{2}, \forall t \in S_m \right\}.
$$

Below, we prove the following lemma.

**Lemma 5** *Under the assumptions of Proposition 5, for $m$ satisfying condition (9), we have*

$$
\mathbb{P}(\Lambda_m^c) \leq c/n^4, \quad \mathbb{P}(\Omega_m^c) \leq c/n^4
$$

*where $c$ is a positive constant.*

Now, we write

$$
\begin{aligned}
\|\widetilde{b}_m - b_A\|_f^2 &= \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m} + \|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c} \\
&= \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} + \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c} + \|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c} (39)
\end{aligned}
$$

From the proof of Theorem 3 in Cohen *et al.* (2013), we get

$$
\mathbb{E}\left( \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} \right) \leq \left( 1 + \frac{8\mathfrak{c}}{\log(n)} \right) \inf_{t \in S_m} (\|t - b_A\|_f^2) + 8\sigma_\varepsilon^2 \frac{m}{n}. \quad (40)
$$

Now we bound the two remaining terms. Clearly, with Lemma 5,

$$
\mathbb{E}(\|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c}) \leq \|b_A\|_f^2 \mathbb{P}(\Lambda_m^c) \leq c/n^4. \quad (41)
$$

Next we deal with $\mathbb{E}(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c})$. We have $\|\hat{b}_m - b_A\|_f^2 \leq 2(\|\hat{b}_m\|_f^2 + \|b_A\|_f^2)$ and

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

First,

$$\|\Psi_m\|_{\rm op} = \sup_{\|\mathbf{x}\|_{2,m}=1} \mathbf{x}'\Psi_{\hat m}\mathbf{x} = \sup_{\|\mathbf{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j\varphi_j(u)\right)^2 f(u)du$$
$$\leq \sup_{\|\mathbf{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j^2 \sum_{j=0}^{m-1} \varphi_j^2(u)\right) f(u)du \leq L(m).$$

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

$$\|\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\|_{\rm op}^2 = \lambda_{\max}\left(\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\widehat{\Phi}_m\widehat{\Psi}_m^{-1}\right) = n\lambda_{\max}(\widehat{\Psi}_m^{-1}) = n\|\widehat{\Psi}_m^{-1}\|_{\rm op}.$$

Therefore, for all $m$ satisfying (9),

$$\|\hat{b}_m\|_f^2 \leq \frac{L(m)\|\widehat{\Psi}_m^{-1}\|_{\rm op}}{n}\left(\sum_{i=1}^n Y_i^2\right) \leq \frac{\mathfrak{c}}{\log(n)}\left(\sum_{i=1}^n Y_i^2\right), \tag{42}$$

and thus on $\Lambda_m$, for $n \geq 3$, $\|\hat{b}_m\|_f^2 \leq C\left(\sum_{i=1}^n Y_i^2\right)$. Then as $\mathbb{E}[(\sum_{i=1}^n Y_i^2)^2] \leq n^2\mathbb{E}(Y_1^4)$, we get

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

On the other hand $\mathbb{E}(\|b_A\|_f^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}) \leq \|b_A\|_f^2\mathbb{P}(\Omega_m^c) \leq c"/n^4$. Thus

$$\mathbb{E}\left(\|\hat{b}_m - b_A\|_f^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}\right) \leq c_1/n. \tag{43}$$

Taking expectation of (39) and plugging (40)-(41)-(43) therein gives the result. □

### 6.6 Proof of Lemma 5

The bound on $\mathbb{P}(\Omega_m^c)$ follows from Proposition 3 under condition (9).

We study now $\mathbb{P}(\Lambda_m^c)$ for $m$ satisfying condition (9). On $\Lambda_m^c$, for $m$ satisfying condition (9), we have $L(m)\|\Psi_m^{-1}\|_{\rm op} \leq \mathfrak{c}n/2\log(n)$ and $L(m)\|\widehat{\Psi}_m^{-1}\|_{\rm op} > \mathfrak{c}n/\log(n)$. This implies, as

$$\mathfrak{c}\frac{n}{\log(n)} < L(m)\|\widehat{\Psi}_m^{-1}\|_{\rm op} \leq L(m)\|\Psi_m^{-1} - \widehat{\Psi}_m^{-1}\|_{\rm op} + L(m)\|\Psi_m^{-1}\|_{\rm op}$$
$$\leq L(m)\|\Psi_m^{-1} - \widehat{\Psi}_m^{-1}\|_{\rm op} + \frac{\mathfrak{c}}{2}\frac{n}{\log(n)},$$

that $L(m)\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\rm op} \geq \mathfrak{c}n/(2\log(n))$. Therefore, we have

$$\Lambda_m^c \subset \{L(m)\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\rm op} > \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}\} \subset \{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\rm op} > \|\Psi_m^{-1}\|_{\rm op}\}.$$

Applying Proposition 4 `(ii)` and Proposition 3, we get

$$\mathbb{P}(\Lambda_m^c) \leq \mathbb{P}\left(\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - {\rm Id}_m\|_{\rm op} \geq \frac{1}{2}\right) \leq \frac{c}{n^4}. \quad \square$$

## 6.7 Proof of Theorem 1

We use the strategy of proof of Theorem 2.11 in Tsybakov (2009). We define proposals $b_0(x) = 0$ and for $\boldsymbol{\theta} = (\theta_0, \dots, \theta_{m-1})'$ with $\theta_j \in \{0, 1\}$,

$$b_\theta(x) = \delta v_n \sigma_\varepsilon \sum_{j=0}^{m-1} \left[\Psi_m^{-1/2}\boldsymbol{\theta}\right]_j \varphi_j(x)$$

where $\Psi_m^{-1/2}$ is a symmetric square-root of the positive definite matrix $\Psi_m^{-1}$.

We choose $v_n^2 = 1/n$ and $m = [n^{1/(s+1)}]$.

• We prove that $b_0, b_\theta \in W_f^s(A, R)$.

As $b_\theta \in S_m$, $(b_\theta)_m^f = b_\theta$ and $(b_\theta)_\ell^f = b_\theta$ for all $\ell \geq m$. Indeed, $S_m \subset S_\ell$. Thus, for $\ell \geq m$, $\|b_\theta - (b_\theta)_\ell^f\|_f^2 = 0$.

Next, $\|b_\theta - (b_\theta)_\ell^f\|_f^2 \leq \|b_\theta\|_f^2$ and as $\int \varphi_j \varphi_k f = [\Psi_m]_{j,k}$, we get

$$\|b_\theta\|_f^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{0 \leq j,k \leq m-1} \left[\Psi_m^{-1/2}\boldsymbol{\theta}\right]_j \left[\Psi_m^{-1/2}\boldsymbol{\theta}\right]_k [\Psi_m]_{j,k} = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{j=0}^{m-1} \theta_j^2 \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m.$$

Thus for $\ell \leq m$,

$$\ell^s \|b_\theta - (b_\theta)_\ell^f\|_f^2 \leq \ell^s \|b_\theta\|_f^2 \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m \ell^s \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m^{s+1} = \delta^2 \sigma_\varepsilon^2.$$

Choosing $\delta$ small enough, we get the result.

• We prove that we can find $\{\theta^{(0)}, \dots, \theta^{(M)}\}$, $M$ elements of $\{0, 1\}^m$ such that $\|b_{\theta^{(j)}} - b_{\theta^{(k)}}\|_f^2 \geq c n^{-s/(s+1)}$ for $0 \leq j < k \leq M$. As above, we find

$$\|b_\theta - b_{\theta'}\|_f^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{j=0}^{m-1} (\theta_j - \theta_j')^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \rho(\theta, \theta'),$$

where $\rho(\theta, \theta') = \sum_{j=0}^{m-1} (\theta_j - \theta_j')^2 = \sum_{j=0}^{m-1} \mathbf{1}_{\theta_j \neq \theta_j'}$ is the Hamming distance between the two binary sequences $\theta$ and $\theta'$. By the Varshamov-Gilbert Lemma (see Lemma 2.9 p.104 in Tsybakov (2009)), for $m \geq 8$, there exists a subset $\{\theta^{(0)}, \dots, \theta^{(M)}\}$ such that $\theta^{(0)} = (0, \dots, 0)$, $\rho(\theta^{(j)}, \theta^{(k)}) \geq m/8$, $0 \leq j < k \leq M$, and $M \geq 2^{m/8}$.

Therefore $\|b_{\theta^{(j)}} - b_{\theta^{(k)}}\|_f^2 \geq \delta^2 v_n^2 \sigma_\varepsilon^2 m/8 = \delta^2 \sigma_\varepsilon^2 n^{-s/(s+1)}/8$.

• Conditional Kullback. Consider first the design $X_1, \dots, X_n$ as fixed. Let $\mathbb{P}_{\theta^{(j)}}^i$ the density of $Y_i = b_{\theta^{(j)}}(X_i) + \varepsilon_i$, i.e. the Gaussian distribution $\mathcal{N}(b_{\theta^{(j)}}(X_i), \sigma_\varepsilon^2)$, and $\mathbb{P}_{\theta^{(j)}}$ the distribution of $(Y_1, \dots, Y_n)$. Then,

$$\frac{1}{M+1} \sum_{j=1}^M K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}}) = \frac{1}{M+1} \sum_{j=1}^M \sum_{i=1}^n \frac{b_{\theta^{(j)}}^2(X_i)}{2\sigma_\varepsilon^2} = \frac{n}{2(M+1)\sigma_\varepsilon^2} \sum_{j=1}^M \|b_{\theta^{(j)}}\|_n^2.$$

Then on $\Omega_n = \cup_{m \leq \mathfrak{c}n/\log(n)} \Omega_m$, we have $\|b_{\theta^{(j)}}\|_n^2 \leq 2\|b_{\theta^{(j)}}\|_f^2$, thus

$$\frac{1}{M+1} \sum_{j=1}^M K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}}) \leq \frac{n\delta^2 v_n^2}{M+1} \sum_{j=1}^M \sum_{k=0}^{m-1} (\theta_k^{(j)})^2 \leq n\delta^2 v_n^2 m \leq \frac{8\delta^2}{\log(2)} \log(M).$$

For $\delta^2$ small enough so that $8\delta^2/\log(2) := \alpha < 1/8$,

$$\frac{1}{M+1}\sum_{j=1}^{M} K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}})\mathbf{1}_{\Omega_n} \leq \alpha \log(M)\mathbf{1}_{\Omega_n}.$$

Now, following Tsybakov (2009), p.116,

$$\begin{aligned}
&\sup_{b_A \in W_f^s(A,R)} \mathbb{E}_{b_A}\left[n^{s/(s+1)}\|T_n - b_A\|_f^2\right] \\
&\geq \mathfrak{A}^2 \max_{b_A \in \{b_{\theta^{(j)}}, j=0,\dots,M\}} \mathbb{P}_{b_A}\left(\|T_n - b_A\|_f > \mathfrak{A}n^{-s/[2(s+1)]}\right) \\
&\geq \mathfrak{A}^2\left(\frac{\log(M+1)-\log(2)}{\log(M)} - \alpha\right)\mathbb{P}(\Omega_n).
\end{aligned}$$

For $n$ large enough and $m$ satisfying (9), it follows from Lemma 5 that $\mathbb{P}(\Omega_n) \geq 1 - (c/n^3) \geq 1/2$. Therefore the lower bound is proved. □

### 6.8 Proof ol Lemma 3 and Proposition 8

For all $\mathbf{u} = (u_0, \dots, u_{m-1})' \in \mathbb{R}^m \setminus \{\mathbf{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j\varphi_j(x)$, $\mathbf{u}'\widehat{\Psi}_m\mathbf{u} = \|t\|_n^2 \geq 0$ . Thus $\|t\|_n = 0 \Rightarrow t(X_i) = 0$ for $i = 1, \dots, n$. As the $X_i$ are almost surely distinct and $t(x)w(x)$ 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 $\mathbf{u} = \mathbf{0}$ and Lemma 3. □

The invertibility of $\Psi_m$ follows from Lemma 1 under (18). Now we prove (23). First note that, for $j$ large enough, $\int \varphi_j^2(x)f(x)dx \leq \frac{c_1}{\sqrt{j}}$, where $c_1$ is a constant. The proof of this Inequality in the Hermite case is given in Belomestny *et al.* (2019), Proposition 2.1. and in Comte and Genon-Catalot (2018) in the Laguerre case. As $\Psi_m$ 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 $\Psi_m$. By (14), we get that for all $j \in \{1, \dots, m\}$, denoting by $\mathbf{e}_j$ the $j$th canonical vector (all coordinates are 0 except the $j$th which is equal to 1), $\mathbf{e_j}'\Psi_m\mathbf{e_j} = \int \varphi_j^2 f$, and

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

As a consequence, $\lambda_{\min}(\Psi_m) \leq c/\sqrt{m}$ which implies the result of Proposition 8. □

### 6.9 Proof of Proposition 9

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

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

We consider the Laguerre functions with index $\delta$, given by

$$\ell_k^{(\delta)}(x) = 2^{(\delta+1)/2}\left(\frac{k!}{\Gamma(k+\delta+1)}\right)^{1/2} L_k^{(\delta)}(2x)e^{-x}x^{\delta/2}, \tag{44}$$

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 $\ell_k^{(\delta)}$, depending on $k$: for $\nu = 4k + 2\delta + 2$, and $k$ large enough, it holds $|\ell_k^{(\delta)}(x/2)| \leq Ce^{-c_0x}$ for $x \geq 3\nu/2$, where $c_0$ is a positive fixed constant.

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 (2018). 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 $n$, odd for odd $n$.

We treat the Laguerre basis first. The result of Askey and Wainger (1965) recalled above states that, for $j$ large enough, $\ell_j(x) \leq ce^{-c_0x}$ for $2x \geq 3(2j+1)$, where $c_02$ is a constant. Thus for $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\|_{2,m} = 1$, we have

$$\mathbf{x}'\Psi_m\mathbf{x} = \int_0^{+\infty}(\sum_{j=0}^{m-1} x_j\ell_j(u))^2 f(u)du \geq \int_0^{3(2m+1)}(\sum_{j=0}^{m-1} x_j\ell_j(\frac{v}{2}))^2 f(\frac{v}{2})\frac{dv}{2}$$

$$\geq \inf_{v\in[0,3(2m+1)]} f(v/2)\int_0^{3(2m+1)/2}(\sum_{j=0}^{m-1} x_j\ell_j(u))^2 du$$

$$\geq \inf_{u\in[0,3(m+1/2)]} f(u)(\int_0^{+\infty}(\sum_{j=0}^{m-1} x_j\ell_j(u))^2 du - \int_{3(m+1/2)}^{+\infty}(\sum_{j=0}^{m-1} x_j\ell_j(u))^2 du).$$

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

$$\int_{3(m+1/2)}^{+\infty}(\sum_{j=0}^{m-1} x_j\ell_j(u))^2 du \leq C'me^{-c_0'm} \leq \frac{1}{2}.$$

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

For the Hermite basis, we proceed analogously using that $|h_j(x)| \leq c|x|e^{-c_0x^2}$ for $x^2 \geq (3/2)(4j+3)$. □

### 6.10 Proof of Inequality (29) of Theorem 2

We denote by $\widehat{M}_n$ the maximal element of $\widehat{\mathcal{M}}_n$ (see (27)) and by $M_n$ the maximal element of $\mathcal{M}_n$ (see (28)). We need also:

$$\mathcal{M}_n^+ = \left\{m \in \mathbb{N}, \quad m\,(\|\Psi_m^{-1}\|_{\text{op}}^2 \vee 1) \leq 4\mathfrak{d}\frac{n}{\log(n)}\right\}, \tag{45}$$

with $\mathfrak{d}$ give in (27). Let $M_n^+$ denote the maximal element of $\mathcal{M}_n^+$. Heuristically, with large probability, considering the constants associated with the sets, we should have $M_n \leq \widehat{M}_n \leq M_n^+$ or equivalently $\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+$, and on this set, we really bound the risk; otherwise, we bound the probability of the complement. More precisely, we denote by

$$\Xi_n := \left\{ \mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+ \right\}, \tag{46}$$

and we write the decomposition:

$$\widehat{b}_{\hat{m}} - b_A = (\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n} + (\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n^c} := T_1 + T_2. \tag{47}$$

The proof relies on two steps and the two following Lemmas.

**Lemma 6** *Under the assumptions of Theorem 2, there exists $\kappa_0$ such that for $\kappa \geq \kappa_0$, we have*

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

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

**Lemma 7** *We have, for $c$ a positive constant, $\mathbb{P}(\Xi_n^c) \leq c/n^2$.*

Lemma 6 gives the bound on $T_1$.
For $T_2$, we use Lemma 7 as follows. Recall that $\Pi_m$ denotes the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\{(t(X_1), \ldots, t(X_n))', t \in S_m\}$ of $\mathbb{R}^n$. We have $\big(\hat{b}_m(X_1), \ldots, \hat{b}_m(X_n)\big)' = \Pi_m Y$. By using the same notation for the function $t$ and the vector $\big(t(X_1), \ldots, t(X_n)\big)'$, we can see that

$$\|b - \hat{b}_{\hat{m}}\|_n^2 = \|b - \Pi_{\hat{m}} b\|_n^2 + \|\Pi_{\hat{m}} \varepsilon\|_n^2 \leq \|b\|_n^2 + n^{-1} \sum_{k=1}^{n} \varepsilon_k^2. \tag{48}$$

Thus

$$\mathbb{E}\left[\|b - \hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n^c}\right] \leq \mathbb{E}\left[\|b\|_n^2 \mathbf{1}_{\Xi_n^c}\right] + \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[\varepsilon_k^2 \mathbf{1}_{\Xi_n^c}] \leq (\mathbb{E}^{1/2}[b^4(X_1)] + \mathbb{E}^{1/2}[\varepsilon_1^4]) \mathbb{P}^{1/2}(\Xi_n^c).$$

We deduce that $\mathbb{E}\left[\|b - \hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n^c}\right] \leq c'/n$. This, together with Lemma 6 plugged in decomposition (47), ends the proof of Inequality (29) of Theorem 2. □

6.11 Proof of Lemma 6.

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

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

Consequently, we can write
$\hat{m} = \arg\min_{m\in\widehat{\mathcal{M}}_n}\{\gamma_n(\hat{b}_m) + \mathrm{pen}(m)\}$, with $\mathrm{pen}(m) = \kappa\sigma_\varepsilon^2\frac{m}{n}$.
Thus, using the definition of the contrast, we have, for any $m \in \widehat{\mathcal{M}}_n$, and any $b_m \in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \le \gamma_n(b_m) + \mathrm{pen}(m). \tag{49}$$

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

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

Thus, on $\Xi_n$, Inequality (49) holds for any $m \in \mathcal{M}_n$ and any $b_m \in S_m$. The decomposition $\gamma_n(t) - \gamma_n(s) = \|t-b\|_n^2 - \|s-b\|_n^2 + 2\nu_n(t-s)$, where $\nu_n(t) = \langle \boldsymbol{\varepsilon}, t\rangle_n$, yields, for any $m \in \mathcal{M}_n$ and any $b_m \in S_m$,

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

We introduce, for $\|t\|_f^2 = \int t^2(u)f(u)du$, the unit ball $B^f_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_f = 1\}$ and the set

$$\Omega_n = \left\{\left|\frac{\|t\|_n^2}{\|t\|_f^2} - 1\right| \le \frac{1}{2},\ \forall t \in \bigcup_{m,m'\in\mathcal{M}_n^+}(S_m + S_{m'}) \setminus \{0\}\right\}. \tag{50}$$

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

$$\begin{aligned}\|\hat{b}_{\hat{m}} - b\|_n^2 \le& \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_{\hat{m}} - b_m\|_f^2 + (8\sup_{t\in B^f_{\hat{m},m}(0,1)}\nu_n^2(t) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}))\\ \le& \left(1+\frac{1}{2}\right)\|b_m - b\|_n^2 + \frac{1}{2}\|\hat{b}_{\hat{m}} - b\|_n^2 + 8\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)}\nu_n^2(t) - p(m,\hat{m})\Big)_+\\ &+ \mathrm{pen}(m) + 8p(m,\hat{m}) - \mathrm{pen}(\hat{m}).\end{aligned} \tag{51}$$

Here we state the following Lemma:

**Lemma 8** *Assume that* **(A1)** *holds, and that* $\mathbb{E}(\varepsilon_1^6) < +\infty$. *Then* $\nu_n(t) = \langle\boldsymbol{\varepsilon}, t\rangle_n$ *satisfies*

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

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

We see that, for $\kappa \geq \kappa_0 = 32$, we have $8p(m, \hat{m}) - \text{pen}(\hat{m}) \leq \text{pen}(m)$. Thus, by taking expectation in (51) and applying Lemma 8, it comes that, for all $m$ in $\mathcal{M}_n$ and $b_m$ in $S_m$,

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n}\big] \leq 3\mathbb{E}\big[\|b_m - b_A\|_n^2\big] + 2\text{pen}(m) + \frac{16\,C}{n}. \tag{52}$$

The complement of $\Omega_n$ satisfies the following Lemma:

**Lemma 9** *Assume that* **(A1)-(A2)** *hold. Then, $\Omega_n$ defined by (50) is such that $\mathbb{P}(\Omega_n^c) \leq c/n^3$ where $c$ is a positive constant.*

We conclude as above (see equation (48)) by writing

$$\mathbb{E}\big[\|b - \hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n^c}\big] \leq \big(\sqrt{\mathbb{E}\big[b^4(X_1)\big]} + \sqrt{\mathbb{E}\big[\varepsilon_1^4\big]}\big)\sqrt{\mathbb{P}(\Omega_n^c)}.$$

This result, together with (52) ends the proof of Lemma 6. □

**Proof of Lemma 8.** We can not apply Talagrand's Inequality to the process $\nu_n$ itself as the noise is not bounded. This is why we decompose the variables $\varepsilon_i$ as follows:

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

Then we have $\nu_n(t) = \nu_{n,1}(t) + \nu_{n,2}(t)$, $\nu_{n,1}(t) = \langle \eta, t\rangle_n$, $\nu_{n,2}(t) = \langle \xi, t\rangle_n$, and

$$\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \leq \Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} 2\nu_{n,1}^2(t) - p(m,\hat{m})\Big)_+ \\ +2 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_{n,2}^2(t). \tag{53}$$

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:

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

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

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

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

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

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

$$\sup_{t\in B_m^f(0,1)} \sup_x |t(x)| \leq c_\varphi\sqrt{m} \sup_{\|\sqrt{\Psi_m}\mathbf{a}\|_{2,m}=1} \|\mathbf{a}\|_{2,m}$$

$$\leq c_\varphi\sqrt{m} \sup_{\|\mathbf{u}\|_{2,m}=1} |\|\sqrt{\Psi_m^{-1}}\mathbf{u}\|_{2,m} = c_\varphi\sqrt{m}\sqrt{\|\Psi_m^{-1}\|_{\mathrm{op}}}.$$

Under condition (45) on $\mathcal{M}_n^+$, we have

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

We can take

$$M_1 := c_\varphi k_n \left(4\mathfrak{d}\frac{n}{\log(n)}\right)^{1/4} (m \vee m')^{1/4}. \tag{54}$$

Consequently, the Talagrand Inequality (see Theorem Klein and Rio (2005) and Theorem A.3 in Supplementary material) implies, for $p(m,m') = 8\sigma_\varepsilon^2 \max(m,m')/n$, and denoting by $m^* := \max(m,m')$,

$$\mathbb{E}\left[\Big(\sup_{t\in B_{m,m'}^f(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+\right] \leq \frac{C_1}{n}(e^{-C_2m^*} + \frac{k_n^2\sqrt{n}(m^*)^{1/2}}{n} e^{-C_3\frac{n^{1/4}(m^*)^{1/4}}{k_n}}).$$

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

$$\mathbb{E}\Big(\sup_{t\in B_{m',m}^f(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+ \leq \frac{C_1'}{n}\left(\exp(-C_2m^*) + (m^*)^{1/2}\exp(-C_3(m^*)^{1/4})\right).$$

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

$$\mathbb{E}[(\sup_{t\in B_{\hat{m},m}^f(0,1)} [\nu_{n,1}]^2(t) - p(m,\hat{m}))_+\mathbf{1}_{\Xi_n}]$$

$$\leq \sum_{m'\in\mathcal{M}_n^+} \mathbb{E}(\sup_{t\in B_{m',m}^f(0,1)} [\nu_{n,1}]^2(t) - p(m,m'))_+ \leq \frac{C}{n}. \tag{55}$$

Let us now study the second term in (53). Recall that $M_n^+ \leq 4\mathfrak{d}n/\log(n)$ the dimension of the largest space of the collection. Then we have

$$\mathbb{E}\Big[\Big(\sup_{t\in B_{\hat{m},m}^f(0,1)} \nu_{n,2}^2(t)\mathbf{1}_{\Xi_n}\Big)_+\Big] \leq \sum_{j=1}^{M_n^+} \mathbb{E}\big[\langle\xi,\bar{\varphi}_j\rangle_n^2\big] = \sum_{j=1}^{M_n^+} \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\bar{\varphi}_j(X_i)\Big)$$

$$= \frac{1}{n}\sum_{j=1}^{M_n^+} \mathbb{E}\big[\xi_1^2\big]\mathbb{E}\big[\bar{\varphi}_j^2(X_1)\big] \leq \frac{M_n^+}{n}\mathbb{E}\big[\varepsilon_1^2\mathbf{1}_{|\varepsilon_1|>k_n}\big] \leq \frac{M_n^+}{n}\frac{\mathbb{E}\big[|\varepsilon_1|^{2+p}\big]}{k_n^p} \leq C\frac{\mathbb{E}\big[\varepsilon_1^6\big]}{n},$$

where the last line follows from the Markov inequality and the choices $k_n = n^{1/4}$ and $p = 4$. This, together with (55) plugged in (53) gives the result. □

**Proof of Lemma 9.** As the collection of models is nested, we have $\mathbb{P}(\Omega_n^c) \le \sum_{m\in\mathcal{M}_n^+} \mathbb{P}(\exists t \in S_m, \left|\frac{\|t\|_n^2}{\|t\|_f^2} - 1\right| > \frac{1}{2}) = \sum_{m\in\mathcal{M}_n^+} \mathbb{P}(\Omega_m^c)$. Now we proved in Lemma 5, that $\mathbb{P}(\Omega_m^c) \le c/n^4$ if $m\|\Psi_m^{-1}\|_{\mathrm{op}} \le (\mathfrak{c}/2)(n/\log(n))$. Here

$$m(\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \le 4\mathfrak{d}\frac{n}{\log(n)} \Rightarrow m\|\Psi_m^{-1}\|_{\mathrm{op}} \le 4\mathfrak{d}\frac{n}{\log(n)}.$$

Therefore, the result holds if $4\mathfrak{d} \le \mathfrak{c}/2$, which is true. With the sum other a set of cardinality less than $n$, we get that $\mathbb{P}(\Omega_n^c) \le c/n^3$. □

## 6.12 Proof of Lemma 7

We study first $\mathbb{P}(\mathcal{M}_n \nsubseteq \widehat{\mathcal{M}}_n) = \mathbb{P}(M_n > \widehat{M}_n)$. On this set, there exists $k \in \mathcal{M}_n$ such that $k \notin \widehat{\mathcal{M}}_n$.

For this index $k$, we have $k\|\Psi_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}n/4\log(n)$ and $k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 > \mathfrak{d}n/\log(n)$. This implies, as

$$\begin{aligned}\mathfrak{d}(n/\log(n)) < k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le 2k \quad &\|\Psi_k^{-1} - \widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 + 2k\|\Psi_k^{-1}\|_{\mathrm{op}}^2 \\ &\le 2k\|\Psi_k^{-1} - \widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 + (\mathfrak{d}/2)(n/\log(n)),\end{aligned}$$

that $k\|\widehat{\Psi}_k^{-1} - \Psi_k^{-1}\|_{\mathrm{op}}^2 \ge \mathfrak{d}n/(4\log(n))$. Let $\Delta_m = \{m\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}}^2 > (\mathfrak{d}/4)n/\log(n)\}$, we have,

$$\mathbb{P}(\mathcal{M}_n \nsubseteq \widehat{\mathcal{M}}_n) \le \sum_{m\in\mathcal{M}_n} \mathbb{P}(\Delta_m) \le \sum_{m\in\mathcal{M}_n} \mathbb{P}(\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \|\Psi_m^{-1}\|_{\mathrm{op}}).$$

We have from `(ii)` of Proposition 4 and Proposition 3, that $\mathbb{P}(\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \|\Psi_m^{-1}\|_{\mathrm{op}}) \le c/n^4$ for $m$ satisfying (9) with $\mathfrak{c}$ given by (10). Indeed, we can conclude as in the proof of Lemma 9 above, because $\mathfrak{d}/4 \le \mathfrak{c}/2$. Thus we proved that $\mathbb{P}(\mathcal{M}_n \nsubseteq \widehat{\mathcal{M}}_n) \le c/n^3$.

Now we study $\mathbb{P}(\widehat{\mathcal{M}}_n \nsubseteq \mathcal{M}_n^+)$. On the set $(\widehat{\mathcal{M}}_n \nsubseteq \mathcal{M}_n^+)$, we can find a $k$ satisfying

$$k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}\frac{n}{\log(n)} \text{ and } k\|\Psi_k^{-1}\|_{\mathrm{op}}^2 > 4\mathfrak{d}\frac{n}{\log(n)},$$

therefore such that $k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}\frac{n}{\log(n)}$ and $k\|\widehat{\Psi}_k^{-1} - \Psi_k^{-1}\|_{\mathrm{op}}^2 \ge \mathfrak{d}\frac{n}{\log(n)}$. Thus we have

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{M}}_n \nsubseteq \mathcal{M}_n^+) &\le \sum_{k\le \mathfrak{d}n/\log(n)} \mathbb{P}(k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}\frac{n}{\log(n)} \text{ and } k\|\widehat{\Psi}_k^{-1} - \Psi_k^{-1}\|_{\mathrm{op}}^2 \ge \mathfrak{d}\frac{n}{\log(n)}) \\ &\le \sum_{k\le \mathfrak{d}n/\log(n)} \mathbb{P}(k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}\frac{n}{\log(n)} \text{ and } \|\widehat{\Psi}_k^{-1} - \Psi_k^{-1}\|_{\mathrm{op}} \ge \|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}).\end{aligned}$$

Now, proceeding with Proposition 4 `(ii)`, interchanging $\widehat{\Psi}_m$ and $\Psi_m$, we get

$$\left\{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\right\} \subset \left\{\|\widehat{\Psi}_m^{-1/2}\Psi_m\widehat{\Psi}_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \frac{1}{2}\right\}.$$

Using $\|\widehat{\Psi}_m^{-1/2}\Psi_m\widehat{\Psi}_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \le \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}}$, we get

$$\left\{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\right\} \subset \left\{\|\widehat{\Psi}_m - \Psi_m\|_{\mathrm{op}} > \frac{1}{2}\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}^{-1}\right\}.$$

Therefore by Proposition 4 and using the value of $\mathfrak{d}$ (this is where $\mathfrak{d}$ is chosen)

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{M}}_n \not\subseteq \mathcal{M}_n^+) &\le \sum_{k\le \mathfrak{d}n/\log(n)} \mathbb{P}(k\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \le \mathfrak{d}\frac{n}{\log(n)} \text{ and } \|\widehat{\Psi}_k - \Psi_k\|_{\mathrm{op}} \ge \frac{1}{2\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}}) \\ &\le \sum_{k\le \mathfrak{d}n/\log(n)} \mathbb{P}(\|\widehat{\Psi}_k - \Psi_k\|_{\mathrm{op}} \ge \frac{1}{2}\sqrt{\frac{k\log(n)}{\mathfrak{d}n}}) \le \frac{c}{n^2}. \square\end{aligned}$$

### 6.13 Proof of Inequality (30) of Theorem 2

We have the following sequence of inequalities, for any $m \in \mathcal{M}_n$ and $t$ any element of $S_m$,

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

where $\Omega_n$ is defined by (50). Therefore, using the result of Theorem 2 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 $t \in S_m$,

$$\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_f^2) \le C_1\left(\|t - b_A\|_f^2 + \sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C_2}{n} + \mathbb{E}\left(\|\hat{b}_{\hat{m}} - b_A\|_f^2\mathbf{1}_{\Omega_n^c}\right), \quad (56)$$

so only the last term is to be studied. First, recall that Lemma 9 implies that $\mathbb{P}(\Omega_n^c) \le \mathfrak{d}/n^3$. Next, write that $\|\hat{b}_{\hat{m}} - b_A\|_f^2 \le 2(\|\hat{b}_{\hat{m}}\|_f^2 + \|b_A\|_f^2)$. As $f$ is bounded, we use a slightly improved version of (42). Indeed, for all $m$,

$$\|\Psi_m\|_{\mathrm{op}} = \sup_{\|\mathbf{x}\|_{2,m}=1} \mathbf{x}'\Psi_m\mathbf{x} = \sup_{\|\mathbf{x}\|_{2,m}=1} \int (\sum_{j=0}^{m-1} x_j\varphi_j(u))^2 f(u)du \le \|f\|_\infty,$$

yields, as for $\hat{m} \in \widehat{\mathcal{M}}_n$, $\|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\mathrm{op}} \vee 1 \le c\sqrt{n}$, $\|\hat{b}_{\hat{m}}\|_f^2 \le \|f\|_\infty\frac{\|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\mathrm{op}}}{n}\left(\sum_{i=1}^n Y_i^2\right) \le \frac{C}{\sqrt{n}}\left(\sum_{i=1}^n Y_i^2\right)$. Then as $\mathbb{E}[(\sum_{i=1}^n Y_i^2)^2] \le n^2\mathbb{E}(Y_1^4)$, we get

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

On the other hand $\mathbb{E}(\|b_A\|_f^2\mathbf{1}_{\Omega_n^c}) \le \|b_A\|_f^2\mathbb{P}(\Omega_n^c) \le c"/n^3$. Thus $\mathbb{E}\left(\|\hat{b}_{\hat{m}} - b_A\|_f^2\mathbf{1}_{\Omega_n^c}\right) \le c_1/n$ and plugging this in (56) ends the proof of Inequality (30) in Theorem 2. □

## References

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.

Askey, R. and Wainger, S. (1965) Mean convergence of expansions in Laguerre and Hermite series. *American Journal of Mathematics* **87**, 695-708.

Baraud, Y. (2000) Model selection for regression on a fixed design. *Probability Theory and Related Fields* **117**, 467-493.

Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probability and Statistics* **6**, 127-146.

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

Belomestny, D., Comte, F., and Genon-Catalot, V. (2016). Nonparametric Laguerre estimation in the multiplicative censoring model. *Electronic Journal of Statistics*, 10(2):3114–3152.

Belomestny, D., Comte, F., and Genon-Catalot, V. (2019). Sobolev-Hermite versus Sobolev nonparametric density estimation on R *The Annals of the Institute of Statistical Mathematics*, **71**, 29-62.

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

Bouaziz, O., Brunel, E. and Comte, F. (2018). Nonparametric survival function estimation for data subject to interval censoring case 2. Preprint hal-01766456.

Brunel, E. and Comte, F. (2009) Cumulative distribution function estimation under interval censoring case 1. *Electrononic Journal of Statistics* **3** , 1-24.

Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics* **13**, 819-834.

Comte, F. and Genon-Catalot, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models. *Electronic Journal of Statistics*, **9**, 1112-1148.

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

Comte, F., Cuenod, C.-A., Pensky, M., and Rozenholc, Y. (2017). Laplace deconvolution and its application to dynamic contrast enhanced imaging. *Journal of the Royal Statistical Society, Series B*, **79**, 69-94.

DeVore, R.A. and Lorentz, G.G. (1993) *Constructive approximation*, Springer-Verlag, Berlin.

Efromovich, S. (1999) *Nonparametric curve estimation. Methods, theory, and applications.* Springer Series in Statistics. Springer-Verlag, New York.

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

Mabon, G. (2017). Adaptive deconvolution on the nonnegative real line. *Scandinavian Journal of Statistics*, 44:707-740.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141-142.

Plancade, S. (2011) Model selection for hazard rate estimation in presence of censoring. *Metrika* **74**, 313-347.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix perturbation theory.* Boston etc.: Academic Press, Inc.

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

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

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

Watson, G.S. (1964) Smooth regression analysis. *Sankhyā, Series A*, **26**, 359-372.