NON PARAMETRIC REGRESSION FUNCTION ESTIMATION IN PRESENCE OF COMMON NOISE.

F. COMTE[(1)], V. GENON-CATALOT[(1)]

Abstract. Consider the regression model $(Y_i = b(X_i) + u_i, i = 1, \dots, n)$ with error process $u_i = \sqrt{\rho}\varepsilon_0 + \sqrt{1-\rho}\varepsilon_i$ where $(\varepsilon_i, i = 0, \dots, n)$ are independent centered random variables (r.v.) with unit variance and $(X_1, \dots, X_n)$ are independent and identically distributed r.v. independent of $(\varepsilon_i, i = 0, \dots, n)$. Thus there is a common noise ε_0 to all u_i's. We study the nonparametric estimation of the regression function b on a subset $A \subset \mathbb{R}$ from the observations $(X_i, Y_i, , i = 1, \dots, n)$ using a projection method on sieves. The standard least-squares contrast fails to provide consistent estimators. Therefore, we introduce a least-squared contrast taking into account the covariance matrix of the noise. By minimizing the contrast over a finite dimensional subspace S_m of $\mathbb{L}^2(A, dx)$, we obtain a projection estimator and study its risk measured either as the expectation of the empirical norm or as the expectation of the theoretical norm associated with the contrast. In the latter case, we study a trimmed estimator. The risk bounds have a variance term with the usual rate. A specific difficulty occurs for proving that the empirical and the theoretical norms are equivalent for functions of S_m. The estimators are implemented on simulated data and show excellent performances when $\mathbb{E}b(X_1)$ is known as this parameter is not identifiable from the model.

May 6, 2025

MSC2010 *Subject classifications.* 62G08 - 62J02
Key words and phrases. Correlated noises, Hermite basis. Nonparametric estimation, Projection estimators, Regression function.

1. Introduction

Consider observations $((X_i, Y_i), i = 1, \dots, n)$ satisfying the regression model

$$Y_i = b(X_i) + u_i, \quad \mathbb{E}u_i = 0, \quad \mathbb{E}u_i^2 = 1, \tag{1}$$

where $(X_1, \dots, X_n)$ are real-valued independent and identically distributed (*i.i.d.*) random variables with common density f, $(u_i, i \geq 1)$ is a stationary sequence of real-valued noises, independent of $(X_1, \dots, X_n)$, the function $b : \mathbb{R} \to \mathbb{R}$ is unknown. It is required to estimate b from the observations $((X_i, Y_i), i = 1, \dots, n)$.
The nonparametric estimation of the regression function b is a very old statistical problem that has been the subject of innumerable contributions especially when (u_i) is a sequence of *i.i.d.* random variables (see *e.g.*, Nadaraya (1964), Watson (1964), Tsybakov (2009) and the references therein). To estimate b, we privilege the sieves method, proposed by Birgé and Massart (1998, 2007), Barron *et al.* (1999), Barron (2002), which provides directly an estimator of b on an

[(1)]: Université Paris Cité, MAP5, UMR 8145 CNRS, FRANCE,
email: valentine.genon-catalot@mi.parisdescartes.fr
.

estimation set $A \subset \mathbb{R}$ and can be briefly described as follows. Choosing a family $(S_m, m \in \mathcal{M}_n)$ of finite-dimensional subspaces of $\mathbb{L}^2(A, dx)$, the standard least-squares contrast

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

is used to build for each m an estimator $b_m^* \in S_m$ by minimizing γ_n^{st} over S_m. Then, an appropriate selection procedure allows to define a data-driven choice of m leading to an adaptive estimator realizing automatically the bias-variance compromise for the risk. The risk of an estimator b^* is measured either as the expectation of the empirical norm $\|b^* - b\|_n^2 = n^{-1}\sum_{i=1}^n (b^*(X_i) - b(X_i))^2$ or as the expectation of $\|b^* - b\|_f^2$ where $\|.\|_f^2$ is the $\mathbb{L}^2(A, f(x)dx)$-norm. The study of the risks is simplified when the estimation set A is assumed to be compact. Nevertheless, recent references have relaxed this constraint (see *e.g.* Comte and Genon-Catalot (2019)). A specific difficulty encountered in the sieves method is to prove the equivalence between the empirical norm $\|.\|_n^2$ and the $\mathbb{L}^2$- norm $\|.\|_f^2$ for functions of S_m.

This programm has been successfully carried out when (u_i) is composed not only of independent variables but also of correlated variables provided that the autocorrelation of the sequence satisfies some constraints. Caron *et al.* (2021) investigate several situations of dependence from short range to long range dependence. In the case of long range dependence, they require that the autocorrelation function of the noise process satisfies $c_u(k) = Corr(u_i, u_{k+i}) \leq \kappa k^{-\gamma}$ for some constant $\kappa > 0$ and $0 < \gamma < 1$. In particular, they need $\gamma > 0$ implying $c_u(k) \to 0$ as k tends to infinity.

In the present paper, we assume that

$$u_i = \sqrt{\rho}\,\varepsilon_0 + \sqrt{1-\rho}\,\varepsilon_i, \quad i = 1, \ldots, n, \quad 0 < \rho < 1, \tag{3}$$

where ρ is known, $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n$ are *i.i.d.* centered with unit variance, *i.e.* there is a common noise ε_0 to all u_i's. It is a way to capture the effect of some random external force influencing simultaneously all individuals such as environment or constraints of energetic or economic type. Such idea has been developped in the theory of mean field games and control for models of stochastic differential equations with common noise (see *e.g.* Carmona *et al.* (2016), Lacker *et al.* (2002), Maillet (2025), Delarue *et al.* (2024)). Indeed, these authors consider a system of stochastic differential equations of the form:

$$dX_i(t) = b(X_i(t))\,dt + \sqrt{\rho}dW_0(t) + \sqrt{1-\rho}dW_i(t), X_i(0) = x, \quad i = 1, \ldots, n, \tag{4}$$

where $W_0, W_1, \ldots, W_n$ are $n+1$ independent Brownian motions. Thus, W_0 is common to all equations and this creates identically distributed but correlated processes. Therefore, this leads to consider the corresponding regression model with a common noise (3) as a toy model in a first step; this turns out to be an interesting question in itself.

Computing the autocorrelation function of (3), we find that $c_u(k) = \rho$ is a fixed quantity. Thus, case (3) is not covered by Caron *et al.* (2021). Up to our knowledge, this kind of autocorrelation function for the noises has not been studied. As a matter of fact, the classical program using (2) fails to provide consistent estimators. Therefore, we consider here a least-squares contrast taking into account the specific dependency, given by:

$$\begin{aligned}\gamma_n(t) &= \frac{1}{n}\left[t(\mathbf{X})^\top R^{-1}\,t(\mathbf{X}) - 2\mathbf{Y}^\top R^{-1}\,t(\mathbf{X})\right] \\ &= \frac{1}{n}\left[(\mathbf{Y} - t(\mathbf{X}))^\top R^{-1}\,(\mathbf{Y} - t(\mathbf{X})) - \mathbf{Y}^\top R^{-1}\mathbf{Y}\right]\end{aligned} \tag{5}$$

where $t(\mathbf{X}) = (t(X_1), \ldots, t(X_n))^\top, \mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $M^\top$ denotes the transpose of any matrix M and R^{-1} is the inverse of the covariance matrix R of $\mathbf{u} = (u_1, \ldots, u_n)^\top$. The contrast (5) is inspired by the exact log-likelihood of $(\mathbf{X}, \mathbf{Y})$. It turns out that R is positive definite for any $\rho \in]0,1[$ and that R^{-1} is simple and explicit, so $\gamma_n(t)$ is explicit when ρ is known. We define the empirical norm:

$$\|t\|_{n,R}^2 := \frac{1}{n} t(\mathbf{X})^\top R^{-1}\, t(\mathbf{X}), \tag{6}$$

together with its theoretical counterpart $\|t\|_R^2 = \mathbb{E}(\|t\|_{n,R}^2)$.
Due to the common noise, the quantity $\mathbb{E}b(X_1)$ is not identifiable in this model. Consequently, we assume that it is known and set $\mathbb{E}b(X_1) = 0$.
In Section 2, we build the minimum contrast estimator $\widehat{b}_m$ by minimizing $\gamma_n(t)$ over a m-dimensional subspace S_m of $\mathbb{L}^2(A, dx)$. In Proposition 1, we give a bound for the risk of the estimator $\widehat{b}_m$ defined as the expectation of the empirical norm (6). We stress that, contrary to the risk bounds in Caron *et al.* (2021), our risk bound has a variance term m/n, similar to one in the case of *i.i.d.* noises. Therefore, the estimator reaches the usual rate of convergence on classical regularity spaces. Section 3 is devoted to the risk defined as the expectation of the theoretical norm. To this end, we introduce a restriction on the dimension of the projection space whichleads to a new stability condition and define a trimmed estimator $\widetilde{b}_m$ where the cut-off corresponds to the empirical version of the stability condition. To study the risk of the trimmed estimator, we introduce a set Ω_m on which the empirical and the theoretical norms defined above are equivalent for functions of S_m. Theorem 1 is devoted to prove that $\mathbb{P}(\Omega_m^c)$ is negligible and Theorem 2 gives the bound for the risk of $\widetilde{b}_m$ measured as the expectation of the theoretical norm. The proof of Theorem 1 is especially long and tough and is obtained via a series of Lemmas. Assuming that ρ is known is a constraint that can be overcome. Indeed, in Section 4, we propose another contrast when ρ is unkown obtained by replacing the covariance matrix which depends on ρ by another well-chosen and known matrix S. This yields another estimator $\widehat{b}_{m,S}$ and its corresponding trimmed version $\widetilde{b}_{m,S}$. The results of the previous sections are extended to the new estimators without any loss of variance (Proposition 5 and Theorem 3). In Section 5, we briefly explain why the standard least-squares contrast fails to provide adequate estimators. Section 6 is devoted to numerical results on simulated data. The method is computationally fast and works in a very convincing way for known $\mathbb{E}(b(X_1))$. An extension if this term is unknown is proposed but requires additional observations. In Section 7, we give some concluding remarks. Section 8 is devoted to proofs. In Section 9, some complementary results are given.

2. Minimum contrast estimator

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 real-valued function h, $h_A = h\mathbf{1}_A$ and we denote $h(\mathbf{X}) = (h(X_1), \dots, h(X_n))^\top$.

2.1. The covariance matrix of the noise.

We have

$$R := \mathrm{Var}(\mathbf{u}) = \Sigma_n(1, \rho) \quad \text{where} \quad \Sigma_n(a,b) = \begin{pmatrix} a & b & \dots & \dots & b \\ b & a & b & \dots & b \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ b & \dots & b & a & b \\ b & \dots & \dots & b & a \end{pmatrix} \in \mathcal{M}_n(\mathbb{R}).$$

Some general properties of the matrices $\Sigma_n(a,b)$ are given in Section 9. The matrix R has 2 eigenvalues: $1-\rho$ with multiplicity $n-1$ and $1+(n-1)\rho$ with multiplicity 1 and associated eigenvector $\mathbf{1}^\top = (1, \dots 1)^\top$. It is positive definite for any $\rho \in]0,1[$ and

$$R^{-1} = \Sigma_n(\alpha_n, \beta_n), \quad \alpha_n = \frac{1+(n-2)\rho}{(1-\rho)(1+(n-1)\rho)}, \quad \beta_n = -\frac{\rho}{(1-\rho)(1+(n-1)\rho)}. \tag{7}$$

The eigenvalues of R^{-1} are the inverse of those of R:

$$\alpha_n - \beta_n = \frac{1}{1-\rho} \quad \text{(multiplicity } n-1) \quad \text{and} \quad \alpha_n + (n-1)\beta_n := c_n = \frac{1}{1+(n-1)\rho}, \tag{8}$$

which has the same eigenvector $\mathbf{1}$. It is easily seen that:

$$R^{-1} = \frac{1}{1-\rho}\left(\mathrm{Id}_n - \frac{\rho}{1+(n-1)\rho}\mathbf{1}\mathbf{1}^\top\right) \tag{9}$$

Moreover, for any $\mathbf{x} = (x_1, \dots, x_n)^\top, \mathbf{y} = (y_1, \dots, y_n)^\top$ vectors of $\mathbb{R}^n$, the following holds:

$$\begin{aligned} \frac{1}{n}\mathbf{x}^\top R^{-1}\mathbf{y} &= \frac{1}{n}\left(\alpha_n \sum_{i=1}^n x_i y_i + \beta_n \sum_{i \neq j} x_i y_j\right) \\ &= \frac{1}{1-\rho}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \frac{n\rho}{1+(n-1)\rho}\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)\right). \end{aligned} \tag{10}$$

2.2. **Identifiability constraint on the regression function.** An intrinsic property of model (2) due to the presence of the common noise ε_0 for all the Y_i's is that the quantity $m = \mathbb{E}(b(X_1))$ is not identifiable. The model can be written

$$\mathbf{Y} = b(\mathbf{X}) + \sqrt{\rho}\varepsilon_0\mathbf{1} + \sqrt{1-\rho}\,\boldsymbol{\varepsilon}.$$

and the standard estimator of m is biased: $\widehat{m}_n = \bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i \to_{a.s.} m + \sqrt{\rho}\varepsilon_0$.
For this reason, in what follows, we will assume that

$$\mathbb{E}[b(X_1)] = 0. \tag{11}$$

We propose in the simulation section a way to overcome this problem.

2.3. **Projection estimator.** Consider model (1) with noises satisfying (3). We assume that $\mathbb{E}\varepsilon_i^4 < \infty$ for $i = 0, 1, \dots, n$ and $\mathbb{E}b^4(X_1) = \int b^4(x)f(x)dx < +\infty$. Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 1, \dots, m)$ be an orthonormal system of A-supported continuous functions belonging to $\mathbb{L}^2(A, dx)$. We assume that for all j, $\|\varphi_j\|_\infty \leq \theta$. Define $S_m = \mathrm{span}(\varphi_1, \dots, \varphi_m)$, the linear space spanned by $(\varphi_1, \dots, \varphi_m)$. We denote by $b_A := b\,\mathbf{1}_A$.

The estimator $\widehat{b}_m$ of b_A on S_m is defined by

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

where $\gamma_n(t)$ is the contrast (5) which takes into account the covariance matrix of the noise $(u_i, i = 1 \dots, n)$. Let us set

$$\widehat{\Psi}_m^{(R)} := \frac{1}{n}\widehat{\Phi}_m^\top R^{-1}\widehat{\Phi}_m, \quad \widehat{\Phi}_m = (\varphi_j(X_i))_{1\leq i\leq n, 1\leq j\leq m} \quad \text{and } \Psi_m^{(R)} := \mathbb{E}[\widehat{\Psi}_m^{(R)}]. \tag{12}$$

To be concrete, for $j, \ell \in \{1, \dots, m\}$, we have

$$[\widehat{\Psi}_m^{(R)}]_{j,\ell} = \frac{1}{n}\varphi_j(\mathbf{X})^T R^{-1}\varphi_\ell(\mathbf{X}).$$

For any function h, we set $\|h\|_{n,R}^2 = n^{-1}h(\mathbf{X})R^{-1}h(\mathbf{X})$ and $\|h\|_R^2 = \mathbb{E}\|h\|_{n,R}^2$. When $t(x) = \sum_{j=1}^m a_j\varphi_j(x) \in S_m$ and $\mathbf{a} = (a_1, \dots, a_m)^\top \in \mathbb{R}^m$, we have

$$\|t\|_{n,R}^2 = \mathbf{a}^T\widehat{\Psi}_m^{(R)}\mathbf{a} \text{ and } \|t\|_R^2 = \mathbf{a}^T\Psi_m^{(R)}\mathbf{a}.$$

Replacing $\mathbf{Y}$ by $b(\mathbf{X}) + \mathbf{u}$, for $\mathbf{u} = (u_1, \dots, u_n)^\top$ yields

$$\gamma_n(t) = \|b - t\|_{n,R}^2 - \|b\|_{n,R}^2 - 2\nu_n(t), \quad \nu_n(t) = \frac{1}{n}\mathbf{u}^\top R^{-1}t(\mathbf{X}).$$

As $\mathbb{E}(\nu_n(t)) = 0$, we obtain $\mathbb{E}[\gamma_n(t)] = \|t - b\|_R^2 - \|b\|_R^2$ which is minimal for $t = b_A$.

By minimizing γ_n over S_m, we obtain that

$$\widehat{b}_m = \sum_{j=1}^m \widehat{a}_j\varphi_j$$

where, provided that $\widehat{\Psi}_m^{(R)}$ is invertible, $\widehat{\mathbf{a}}_m = (\widehat{a}_1, \dots, \widehat{a}_m)^\top$ is given by

$$\widehat{\mathbf{a}}_m = (\widehat{\Psi}_m^{(R)})^{-1}\widehat{Z}_m, \quad \widehat{Z}_m = \frac{1}{n}\widehat{\Phi}_m^\top R^{-1}\mathbf{Y}. \tag{13}$$

More precisely, for $j \in \{1, \dots, m\}$, $[\widehat{Z}_m]_j = (1/n)\varphi_j(\mathbf{X})^T R^{-1}\mathbf{Y}$. Note that

$$\begin{aligned} \widehat{Z}_m &= \frac{1}{n}\widehat{\Phi}_m^T R^{-1}b(\mathbf{X}) + \frac{1}{n}\widehat{\Phi}_m^T R^{-1}\mathbf{u} \\ &= \frac{1}{n}\widehat{\Phi}_m^T R^{-1}b(\mathbf{X}) + \frac{\sqrt{1-\rho}}{n}\widehat{\Phi}_m^T R^{-1}\boldsymbol{\varepsilon} + \sqrt{\rho}\varepsilon_0\frac{c_n}{n}\widehat{\Phi}_m^T\mathbf{1}, \end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ and $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$, $\mathbf{u} = \sqrt{\rho}\varepsilon_0\mathbf{1} + \sqrt{1-\rho}\,\boldsymbol{\varepsilon}$ and we recall that $R^{-1}\mathbf{1} = c_n\mathbf{1}$. The first two terms of $\widehat{Z}_m$ correspond to standard nonparametric regression when $\rho = 0$ (so $R^{-1} = \mathrm{Id}_n$). If n is large, R^{-1} tends to Id_n and c_n tends to 0.

2.4. **Risk bound as expectation of the empirical norm.** We can prove the following bound.

Proposition 1. *Assume that $\widehat{\Psi}_m^{(R)}$ is invertible. The risk bound of $\widehat{b}_m$ expressed as the expectation of the empirical norm $\|.\|_{n,R}$ satisfies*

$$\mathbb{E}[\|\widehat{b}_m - b_A\|_{n,R}^2] \leq \inf_{t \in S_m} \|t - b_A\|_R^2 + \frac{m}{n}.$$

It is noteworthy that in spite of the common noise, we obtain a risk bound similar to the usual case with a preserved variance term m/n. For expliciting the bias, we state a simple Lemma.

Lemma 1. *The following properties hold:*

(1) $\|t\|_{n,R}^2 \geq 0$ *with*

$$\frac{1}{1+(n-1)\rho}\|t\|_n^2 \leq \|t\|_{n,R}^2 \leq \frac{1}{1-\rho}\|t\|_n^2$$

where $\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n t^2(X_i)$.

(2) $\|t\|_R^2 \geq 0$ *with*

$$\frac{1}{1+(n-1)\rho}\mathbb{E}[t^2(X_1)] \leq \|t\|_R^2 \leq \frac{1}{1-\rho}\mathbb{E}[t^2(X_1)].$$

Obviously, Lemma 1 implies

$$\inf_{t\in S_m}\|t-b_A\|_R^2 \le \frac{1}{1-\rho}\inf_{t\in S_m}\|t-b_A\|_f^2$$

If in addition, the density f of X_1 is bounded, we get

$$\inf_{t\in S_m}\|t-b_A\|_R^2 \le \frac{\|f\|_\infty}{1-\rho}\|b_m-b_A\|^2$$

where b_m is the orthogonal projection of b_A on S_m. Thus, we can make the standard bias variance compromise and obtain usual nonparametric rates for the risk bound of Proposition 1.

3. Risk bound expressed as expectation of the theoretical norm

The aim of this section is to handle the integrated risk with respect to the squared norm $\|.\|_R^2$. It appears that this task is really far from simple is the context of regression with common noise.

3.1. First properties of the empirical and the theoretical norms.

Set

$$\Psi_m = (\mathrm{cov}(\varphi_j(X_1),\varphi_k(X_1))_{1\le j,k\le m}\,. \tag{14}$$

Let us stress that in the classical regression ($\rho = 0$), the reference matrix is not a matrix of covariances, but is equal to $(\mathbb{E}(\varphi_j(X_1)\varphi_k(X_1)))_{1\le j,k\le m}$, see Baraud (2000), Cohen *et al.* (2013, 2019), Comte and Genon-Catalot (2019) and references therein.

Proposition 2.

- *For the theoretical norm, the following decomposition holds.*

$$\|t\|_R^2 = \alpha_n \mathrm{Var}(t(X_1)) + c_n\{\mathbb{E}[t(X_1)]\}^2,$$

where $c_n = \alpha_n + (n-1)\beta_n = \frac{1}{1+(n-1)\rho}$.

For $t(x) = \sum_{j=1}^m a_j\varphi_j(x)$,

$$\|t\|_R^2 = \ \mathbf{a}^T\Psi_m^{(R)}\mathbf{a} = \alpha_n\ \mathbf{a}^T\Psi_m\mathbf{a} + c_n\mathbf{a}^T\Xi_m\mathbf{a} \tag{15}$$

where $\Xi_m = (\mathbb{E}\varphi_j(X_1)\mathbb{E}\varphi_k(X_1))_{1\le j,k\le m}$ *and* Ψ_m *is defined by (14).*

- *For the empirical norm, the following decomposition holds.*

$$\|t\|_{n,R}^2 = \alpha_n Z_n(t) + c_n\{\mathbb{E}[t(X_1)]\}^2 + 2c_n\mathbb{E}[t(X_1)]V_n(t) + (n-1)\beta_n U_n(t), \tag{16}$$

with

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^n[t(X_i)-\mathbb{E}(t(X_i))]^2, \quad V_n(t) = \frac{1}{n}\sum_{i=1}^n[t(X_i)-\mathbb{E}(t(X_i))] \tag{17}$$

$$U_n(t,s) = \frac{1}{n(n-1)}\sum_{1\le i\ne k\le n}[t(X_i)-\mathbb{E}(t(X_i))][s(X_k)-\mathbb{E}(s(X_k))], \quad U_n(t) = U_n(t,t). \tag{18}$$

Moreover

$$\|t\|_{n,R}^2 = \mathbf{a}^T\widehat{\Psi}_m^{(R)}\mathbf{a} \to_{n\to\infty} \frac{1}{1-\rho}\mathbf{a}^T\Psi_m\mathbf{a},\ \textit{almost surely.}$$

Note that $\|t\|_R^2$ and $\Psi_m^{(R)}$ are deterministic but still depend on n. However, for any function t,

$$\lim_{n\to+\infty}\|t\|_R^2 = \frac{1}{1-\rho}\mathrm{Var}[t(X_1)] = \frac{1}{1-\rho}\,\mathbf{a}^T\Psi_m\mathbf{a}.$$

From Proposition 2, we have $\|t\|_R^2 \ge 0$. If $\|t\|_R^2 = 0$, then $\mathrm{Var}(t(X_1)) = 0$ and $\mathbb{E}(t(X_1)) = 0$. This implies that $\mathbb{E}[t^2(X_1)] = 0$, which in turn implies that $t(x) = 0$ almost surely on the support of the distribution of X_1. If the support of X_1 contains an interval and if $t \in S_m$, as the basis functions are continuous, $t = 0$.

If one of the basis functions, say φ_1 is constant, then $\mathrm{cov}(\varphi_1(X_1), \varphi_j(X_1)) = 0$ for $j = 1, \dots, m$. In such a case, Ψ_m is not inversible. Therefore, it is mandatory to choose a basis which does not contain any constant function. This is why we exclude the trigonometric basis and choose in Section 6 the Hermite basis which does not contain any constant function.

Moreover, we have

$$\|t\|^2_{n,R} = \frac{1}{1-\rho}\frac{1}{n}\sum_{i=1}^n \left[t(X_i) - \frac{1}{n}\sum_{i=1}^n t(X_i)\right]^2 + c_n\left(\frac{1}{n}\sum_{i=1}^n t(X_i)\right)^2.$$

and therefore $\|t\|_{n,R} = 0$ implies $t(X_i) = 0$ a.s., for $i = 1, \dots, n$. If $t \in S_m$, $n \geq m$ and there exists a m-sub-sample $\{X_{i_1}, \dots, X_{i_m}\} \subset \{X_1, \dots, X_n\}$ such that $\det((\varphi_j(X_{i_k}))_{1\leq j,k\leq m}) \neq 0$, then $t = 0$. In that case, $\widehat{\Psi}^{(R)}_m$ is invertible and the estimator is well-defined.

We stress that decomposition (16)-(18) is crucial for the comparison of $\widehat{\Psi}^{(R)}_m$ and $\Psi^{(R)}_m$.

3.2. **Equivalence of norms and trimmed estimator.**

We assume that

$$L(m) := \sup_{x\in\mathbb{R}} \sum_{j=0}^{m-1} \varphi_j^2(x) < +\infty \tag{19}$$

and that $L(m) \geq 1$ (otherwise it is possible to change $L(m)$ into $L(m) \vee 1$ everywhere). In most classical bases, $L(m)$ is of order m, see *e.g.* Comte and Genon-Catalot (2019); for the Hermite basis, $L(m)$ is of order $\sqrt{m}$, see Lemma 1 in Comte and Lacour (2023)

As proved in Cohen *et al.* (2013, 2019), in order to ensure stability of a least-squares estimator, a restriction on the possible dimensions of the projection space is mandatory. In our context, the stability condition takes the following form. We consider m such that

$$L(m)(\|(\Psi^{(R)}_m)^{-1}\|_{\mathrm{op}} \vee 1) \leq c^\star \frac{n}{\log(n)}, \tag{20}$$

where $c^\star$ is a constant which will be precised later on.

Note that by Lemma 1 (2), the eigenvalues of $\Psi^{(R)}_m$ belong to $\left[\frac{1}{1+(n-1)\rho}, \frac{1}{1-\rho}\right]$. Therefore, $\Psi^{(R)}_m$ is invertible and the eigenvalues of $(\Psi^{(R)}_m)^{-1}$ belong to $[1-\rho, 1+(n-1)\rho]$. Consequently, condition (20) is a true restriction. Moreover, although $\Psi^{(R)}_m$ is invertible, Ψ_m may be non invertible. Nevertheless, exploiting the link between the matrix $\Psi^{(R)}_m$ and the matrix Ψ_m defined by (14), we can prove that the stability condition (20) implies that Ψ_m is invertible and that moreover the following holds:

Proposition 3. *Under condition (20), for* $n \geq n_0 = \exp\left(\frac{2c^\star\|f\|^2}{\rho}\right) \vee 3$, *it holds*

$$L(m)\|\Psi_m^{-1}\|_{\mathrm{op}} \leq 4c^\star(1-\rho)\frac{n}{\log(n)}.$$

We now introduce a cutoff for the estimator, which represents the empirical version of the stability condition (20). We set

$$\widetilde{b}_m = \widehat{b}_m \mathbf{1}_{\Lambda_m}, \quad \Lambda_m = \left\{L(m)(\|(\widehat{\Psi}^{(R)}_m)^{-1}\|_{\mathrm{op}} \vee 1) \leq kc^\star\frac{n}{\log(n)}\right\} \tag{21}$$

where k is an integer to be chosen below (namely $k = 4$, see Proposition 4).

The classical method to study the integrated risk of the trimmed estimator, is to define a set Ω_m on which the empirical norm $\|t\|^2_{n,R}$ and the deterministic norm $\|t\|^2_R$ are equivalent for all functions of S_m and such that Ω_m^c has small probability (see *e.g.* Cohen et al. (2013, 2019), Baraud (2002), Comte and Genon-Catalot (2019)). This part is especially difficult in our context. Let us define

$$\Omega_m = \left\{ \forall t \in S_m, t \neq 0, \left| \frac{\|t\|_{n,R}^2}{\|t\|_R^2} - 1 \right| \leq \frac{3}{4} \right\}. \tag{22}$$

Proposition 4. *We have* $\Omega_m = \left\{ \|((\Psi_m^{(R)})^{-1/2} \widehat{\Psi}_m^{(R)} (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leq \frac{3}{4} \right\}$ *and under condition (20), by setting $k = 4$ in the definition of Λ_m given in (21), we have:* $\Lambda_m^c \subset \Omega_m^c$.

Theorem 1 below relies on several steps and decomposition (16) is a key tool to prove it.

Theorem 1. *Recall that the basis functions are bounded by θ and take $m \leq c^\dagger n/\log(n)$ for some constant $c^\dagger > 0$. Assume that condition (20) holds with*

$$c^\star = \frac{1}{4(1-\rho)} \left(\mathbf{c}_p(\frac{1}{2}) \wedge \frac{1}{16\mathbf{c}_1} \right), \; c_p(\frac{1}{2}) = \frac{1 - 3\log(3/2)}{8(p+1)}, \; \mathbf{c}_1 = 48 c_0 (\theta^2 \vee 1)(c^\dagger \vee 1)(p+2)^2, \tag{23}$$

where c_0 is a numerical constant ($c_0 = 62.3$). Then, for

$$n \geq n_0 := \exp\left(\frac{2c^\star \|f\|^2}{\rho} \right) \vee \frac{192}{\rho}, \tag{24}$$

it holds that $\mathbb{P}(\Omega_m^c) \leq C n^{-p}$.

The sketch of the proof is as follows. For proving that $\mathbb{P}(\Omega_m)^c \lesssim n^{-p}$, we introduce two sets $\Omega_m^{(1)}$ and $\Omega_m^{(2)}$. We take $\Omega_m^{(1)} = \Omega_m^{(1)}(1/2)$ where for $0 < \delta < 1$,

$$\Omega_m^{(1)}(\delta) := \left\{ \sup_{t \in S_m, \mathrm{Var}(t(X_1)) \neq 0} \left| \frac{Z_n(t)}{\mathrm{Var}(t(X_1))} - 1 \right| \leq \delta \right). \tag{25}$$

We also define, for c_0 is a numerical constant ($c_0 = 62.3$) and $m \leq c^\dagger n/\log(n)$,

$$\Omega_m^{(2)} := \left\{ \sup_{t \in S_m, \|t\| \neq 0} \left| \frac{U_n(t,t)}{\|t\|^2} \right| \leq 24 c_0 (\theta^2 \vee 1)\,(p+2)^2 (L(m) + c^\dagger) \frac{\log(n)}{n} \right\} \tag{26}$$

(see (16)). We prove that for $n \geq n_0$, $\mathbb{P}(\Omega_m^{(1)})^c \lesssim n^{-p}$ (Lemma 3) and $\mathbb{P}(\Omega_m^{(2)})^c \lesssim n^{-p}$ (Lemma 4). We conclude by proving that $\Omega_m^{(1)} \cap \Omega_m^{(2)} \subset \Omega_m$ hence $\mathbb{P}(\Omega_m^c) \lesssim n^{-p}$.

On Ω_m, the empirical and the theoretical norms are equivalent for functions of S_m. Now, we can state:

Theorem 2. *Assume that $m \leq c^\dagger n/\log(n)$ and that condition (20) holds with $c^\star$ defined in (23) below with $p \geq 6$. Then for $n \geq n_0$ (see (24)), it holds that*

$$\mathbb{E}[\|\widetilde{b}_m - b_A\|_R^2] \leq C_1 \left(\inf_{t \in S_m} \|t - b_A\|_R^2 + \frac{m}{n} \right) + \frac{C_2}{n}$$

where C_1 is a numerical positive constant ($C_1 = 18$ suits) and $C_2 > 0$ depends on $\|f\|_\infty, \rho$.

4. Another contrast if ρ is unknown

Now, we replace R^{-1} by $S^{-1} = \Sigma_n(1, -\frac{1}{n})$ and consider the contrast

$$\Gamma_n(t) = \frac{1}{n} \left[t(\mathbf{X})^\top S^{-1} t(\mathbf{X}) - 2\mathbf{Y}^\top S^{-1} t(\mathbf{X}) \right], \tag{27}$$

The matrix S^{-1} is positive definite with eigenvalues $1+\frac{1}{n}$ (with multiplicity $n-1$) and $1-\frac{n-1}{n}=\frac{1}{n}$. We define

$$\|t\|_{n,S}^2 = t(\mathbf{X})^\top S^{-1}\, t(\mathbf{X}), \quad \|t\|_S^2 = \mathbb{E}\|t\|_{n,S}^2, \quad \nu_{n,S}(t) = \frac{1}{n}\mathbf{u}^\top S^{-1}\, t(\mathbf{X}).$$

Note that $S=\Sigma(2n/(n+1), n/(n+1))$. All computations of Lemma 2 hold provided (α_n,β_n) is replaced by $(1,-\frac{1}{n})$ and $(\frac{1}{1-\rho}, \frac{1}{1+(n-1)\rho})$ is replaced by $(1+\frac{1}{n},\frac{1}{n})$.

When minimizing Γ_n over S_m, we obtain $\widehat{b}_{m,S}=\sum_{j=1}^m \widehat{a}_j^S\varphi_j$ with $\widehat{\mathbf{a}}_m^S=(\widehat{a}_1^S,\ldots,\widehat{a}_m^S)^\top$:

$$\widehat{\mathbf{a}}_m^S = \left(\widehat{\Phi}_m^\top S^{-1}\,\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top S^{-1}\mathbf{Y} = (\widehat{\Psi}_m^{(S)})^{-1}\left(\frac{1}{n}\widehat{\Phi}_m^\top S^{-1}\mathbf{Y}\right),$$

where as previously, $\widehat{\Phi}_m=(\varphi_j(X_i))_{1\le i\le n, 1\le j\le m}\in\mathcal{M}_{n,m}(\mathbf{R})$, and here

$$\widehat{\Psi}_m^{(S)} = \widehat{\Phi}_m^\top S^{-1}\,\widehat{\Phi}_m, \quad \Psi_m^{(S)} = \mathbb{E}[\widehat{\Psi}_m^{(S)}].$$

We can prove the following result:

Proposition 5. *In model* (1)*, the estimator* $\widehat{b}_{m,S}$ *satisfies:*

$$\mathbb{E}[\|\widehat{b}_{m,S}-b_A\|_{n,S}^2] \le \inf_{t\in S_m}\|t-b_A\|_S^2 + \frac{2m}{n}.$$

For the integrated S-norm, we define the stability condition:

$$L(m)(\|(\Psi_m^{(S)})^{-1}\|_{\mathrm{op}}\vee 1) \le \mathbf{c}_S^\star\frac{n}{\log(n)}, \quad \mathbf{c}_S^\star = \frac{1}{2}\inf\left\{\mathbf{c}_p(1/2), \frac{1}{8\mathbf{c}_1}\right\}, \tag{28}$$

where, $\mathbf{c}_p(1/2)$, $\mathbf{c}_1$ are given by (23), and by convention, $\|(\Psi_m^{(S)})^{-1}\|_{\mathrm{op}}=+\infty$ whenever $\Psi_m^{(S)}$ is not invertible. We introduce a cutoff for the estimator and set

$$\widetilde{b}_{m,S}=\widehat{b}_{m,S}\mathbf{1}_{\Lambda_{m,S}}, \quad \Lambda_{m,S}=\left\{L(m)(\|(\widehat{\Psi}_m^{(S)})^{-1}\|_{\mathrm{op}}\vee 1)\le 4c_S^\star\frac{n}{\log(n)}\right\}. \tag{29}$$

Proposition 3 and Theorem 3 are extended as follows.

Lemma 2. *Under condition (28), for* $n\ge n_0=\exp\left(2c^\star\|f\|^2\right)$*, it holds*

$$L(m)\|\Psi_m^{-1}\|_{\mathrm{op}} \le 2c_S^\star\frac{n}{\log(n)}.$$

Theorem 3. *Assume that* $m\le c^\dagger n/\log(n)$ *and that condition (20) holds with* $c^\star$ *defined in (23) with* $p\ge 6$*. Then for* $n\ge \exp(2c_S^\star\|f\|^2)\vee 192$*, it holds that*

$$\mathbb{E}[\|\widetilde{b}_{m,S}-b_A\|_S^2] \le C_1'\left(\inf_{t\in S_m}\|t-b_A\|_S^2+\frac{m}{n}\right)+\frac{C_2'}{n}$$

where C_1' *is a numerical positive constant* ($C_1'=18$ *suits) and* $C_2'>0$ *depends on* $\|f\|_\infty$.

5. Comparison with the standard least squares

In this paragraph, we consider the standard contrast γ_n^{st} (see (2)) and look at the corresponding minimum contrast estimators. The empirical and theoretical norms corresponding to γ_n^{st} are

$$\|t\|_n^2 = n^{-1}t(\mathbf{X})^\top t(\mathbf{X}), \quad \mathbb{E}\|t\|_n^2 = \|t\|_f^2.$$

For $t(x)=\sum_{j=1}^m a_j\varphi_j(x)$, $\|t\|_n^2 = a^\top\widehat{\Psi}_m^{st}a$ with $\widehat{\Psi}_m^{st}=n^{-1}\widehat{\Phi}_m^\top\widehat{\Phi}_m$ and

$$\mathbb{E}[\widehat{\Psi}_m^{st}] = \Psi_m^{st} = (\mathbb{E}(\varphi_j(X_1)\varphi_k(X_1))_{1\le j,k\le m}.$$

We can write $\gamma_n(t) = \|b-t\|_n^2 - \|b\|_n^2 - 2\nu_n^{st}(t)$ where

$$\nu_n^{st}(t) = n^{-1}\mathbf{u}^\top t(\mathbf{X}) = n^{-1}(\sqrt{\rho}\,\varepsilon_0 \mathbf{1}^\top t(\mathbf{X}) + \sqrt{1-\rho}\,\boldsymbol{\varepsilon}^\top t(\mathbf{X})).$$

We have $\mathbb{E}\nu_n^{st}(t) = 0$ and some easy computations lead to

$$\mathbb{E}[\nu_n^{st}(t)]^2 = \frac{1}{n}\mathrm{Var}(t(X_1)) + \rho[\mathbb{E}(t(X_1)]^2.$$

The variance of $\nu_n^{st}(t)$ does not tend to 0 unless $\mathbb{E}(t(X_1)) = 0$.

Define $\widehat{b}_m^{st}$ as the minimizer of γ_n^{st} on S_m. To study this estimator, we simply replace R^{-1} by Id_n and c_n by 1. Mimicking the proof of Proposition 1 with $R^{-1} = \mathrm{Id}_n$ and $c_n = 1$, we have $\widehat{b}_m^{st} = \sum_{j=1}^m \widehat{a}_j^{st}\varphi_j$ with

$$\widehat{mathbf a}_m^{st} = (\widehat{\Psi}_m^{st}) - 1\widehat{Z}_m^{st} \quad \text{where} \quad \widehat{Z}_m^{st} = \frac{1}{n}\widehat{\Phi}_m^\top \mathbf{Y}.$$

With $Z_m^{st} = \frac{1}{n}\widehat{\Phi}_m^\top b(\mathbf{X})$, we have $\widehat{Z}_m^{st} = Z_m^{st} + \sqrt{\rho}\varepsilon_0\frac{1}{n}\widehat{\Phi}_m^\top\mathbf{1} + \sqrt{1-\rho}\frac{1}{n}\widehat{\Phi}_m^\top\boldsymbol{\varepsilon}$ and

$$\|b - \widehat{b}_m^{st}\|_n^2 = \inf_{t\in S_m}\|b-t\|_n^2 + V_m(n), \quad \text{where} \tag{30}$$

$$V_m(n) := (\widehat{Z}_m^{st} - Z_m^{st})^\top(\widehat{\Psi}_m^{st})^{-1}(\widehat{Z}_m^{st} - Z_m^{st}) = \mathbf{u}^\top\frac{1}{n}\widehat{\Phi}_m(\frac{1}{n}\widehat{\Phi}_m^\top\widehat{\Phi}_m)^{-1}\frac{1}{n}\widehat{\Phi}_m^\top\mathbf{u}$$

Thus,

$$\mathbb{E}V_m(n) = \mathbb{E}\mathrm{Tr}[\frac{1}{n}\widehat{\Phi}_m(\widehat{\Psi}_m^{st})^{-1}\frac{1}{n}\widehat{\Phi}_m^\top\mathbf{u}\mathbf{u}^\top]$$

where $\mathbb{E}\mathbf{u}\mathbf{u}^\top = (1-\rho)\mathrm{Id}_n + \rho\mathbf{1}\mathbf{1}^\top$. Consequently,

$$\mathbb{E}V_m(n) = (1-\rho)\frac{m}{n} + \rho\mathbb{E}B_m(n), \quad \text{where} \quad B_m(n) = n^{-1}\mathbf{1}^\top\widehat{\Phi}_m(n^{-1}\widehat{\Phi}_m^\top\widehat{\Phi}_m)^{-1}n^{-1}\widehat{\Phi}_m^\top\mathbf{1}.$$

Clearly, $B_m(n)$ tends as n tends to infinity to a fixed limit equal to

$$B_m = \mathbb{E}[\boldsymbol{\varphi}(X_1)^\top]\,(\Psi_m^{st})^{-1}\mathbb{E}[\boldsymbol{\varphi}(X_1)] > 0,$$

where $\boldsymbol{\varphi}(X_1)^\top = (\varphi_1(X_1), \dots, \varphi_m(X_1))$. It holds that $B_m > 0$ if $\exists j_0 \in \{1, \dots, m\}$ such that $\mathbb{E}[\varphi_{j_0}(X_1)] \neq 0$; in particular $B_m > 0$ in all the examples of Section 6. Therefore, looking at (30), we see that the risk bound contains an additional term which is not present in the standard case of *i.i.d.* noises.

6. Numerical simulation results

In this section, we consider first the case where $\mathbb{E}[b(X_1)]$ is known and choose functions and the density f such that it is nul. Then, we consider additional observations leading to a consistent estimation of $\mathbb{E}[b(X_1)]$ that we use in the estimation procedure based on n trajectories. The basis functions is the Hermite basis (see *e.g.* Comte and Genon-Catalot, 2018).

		$\rho = 0$				$\rho = 0.25$						$\rho = 0.75$					
		$n = 250$		$n = 1000$		$n = 250$			$n = 1000$			$n = 250$			$n = 1000$		
		R	O	R	O	R	S	O	R	S	O	R	S	O	R	S	O
	MSE	6.42	4.97	1.73	1.33	7.06	7.16	6.02	1.84	1.88	1.54	4.00	4.12	3.55	1.09	1.32	1.02
b_1	std	3.54	2.82	0.94	0.75	4.45	4.72	4.12	1.19	1.20	1.06	3.63	3.51	3.46	1.08	1.10	1.07
	dim	8.06	9.19	9.13	10.2	8.27	8.03	9.28	9.37	9.01	10.3	8.79	7.89	10.1	10.1	8.65	11.0
	MSE	10.0	8.36	2.76	2.17	11.0	11.0	9.63	2.97	3.00	2.45	6.08	6.17	5.53	1.55	1.94	1.45
b_2	std	5.56	4.46	1.37	1.10	7.00	6.98	6.64	1.93	1.95	1.70	5.67	5.66	5.53	1.44	1.47	1.38
	dim	8.29	9.11	8.91	9.91	8.35	8.16	9.31	9.23	8.82	10.2	8.84	8.01	9.97	10.0	8.84	10.8
	MSE	15.0	10.3	3.73	3.02	13.3	14.6	10.7	3.58	3.36	3.01	6.87	7.16	6.19	1.97	1.83	1.68
b_3	std	12.0	6.33	2.29	1.81	10.2	12.0	7.19	2.16	2.13	1.86	5.59	7.43	5.43	1.46	1.44	1.35
	dim	5.95	6.53	6.50	7.23	6.24	5.86	6.77	6.62	6.24	7.31	6.39	5.91	7.33	7.10	6.01	8.49
	MSE	5.06	3.98	1.28	1.07	5.75	5.95	4.82	1.47	1.48	1.28	3.29	3.79	3.03	0.86	0.87	0.80
b_4	std	2.54	2.01	0.67	0.53	3.72	3.84	3.23	0.99	1.01	0.90	3.18	3.21	3.11	0.90	0.91	0.90
	dim	8.84	9.85	8.84	9.95	9.03	8.80	10.0	10.2	10.1	11.2	9.63	8.46	10.9	10.3	9.96	11.9

TABLE 1. Results for beta noise with 400 repetitions, for three values of ρ ($\rho = 0, 0.25, 0.75$), two sample sizes (n=250, 1000), and the method with R^{-1} ('R'), the one with S^{-1} ('S'), compared to the oracle ('O'), for functions defined in (31). MSE is 1000× Relative MISE, std is 1000×std, and 'dim' is the mean of the selected dimensions.

6.1. **Case $\mathbb{E}[b(X_1)]$ is known.** If $\mathbb{E}[b(X_1)]$ is known, then the variables Y_i are replaced by the $Y_i - \mathbb{E}[b(X_1)]$, the function estimated is thus $b(X_1) - \mathbb{E}[b(X_1)]$, and the constant can be finally added to the estimate to recover the right level for b. This is why we consider hereafter examples where $\mathbb{E}[b(X_1)] = 0$.

We consider four examples of odd functions

$$b_1(x) = 2x, \quad b_2(x) = 3\sin(0.8\,\pi x), \quad b_3(x) = \frac{4x}{1+x^2}, \quad b_4(x) = 1.75x^3\exp(-0.5|x|) \tag{31}$$

where the multiplicative constants are such that the empirical standard deviation of $b_i(\mathbf{X})$ is of order 2 (or slightly less), as a measure of the signal to noise ratio (the noise has unit variance). The X_i's follow a symmetric distribution so that $\mathbb{E}[b_i(X_1)] = 0$: either $6 \times (\beta(3,3) - 0.5)$ or a $\mathcal{N}(0,1)$. The ε_i are Gaussian $\mathcal{N}(0,1)$.

We consider the estimator based on matrix R, abbreviated as the 'R-procedure' in the sequel. We propose a model selection procedure to get a data-driven choice $\widehat{m}$ of the dimension m. The selection of m is performed among dimensions $m = 1, \dots, 20$, such that the constraint defined for Λ_m in (21) is fulfilled, that is

$$\widehat{\mathcal{M}}_n^{(R)} = \left\{ m \in \{1, \dots, 20\},\ \sqrt{m}(\|(\widehat{\Psi}_m^{(R)})^{-1}\|_{\mathrm{op}} \vee 1) \le \frac{c_0}{1-\rho}\frac{n}{\log(n)} \right\}, \quad c_0 = 8 \times 64\frac{n}{\log(n)}.$$

Then, we apply the classical procedure

$$\widehat{m}^{(R)} = \arg\min_{m \in \widehat{\mathcal{M}}_n^{(R)}} \left\{ -\|\widehat{b}_m\|_{n,R}^2 + \mathrm{pen}(m) \right\}, \quad \mathrm{pen}(m) = \kappa\frac{m}{n},$$

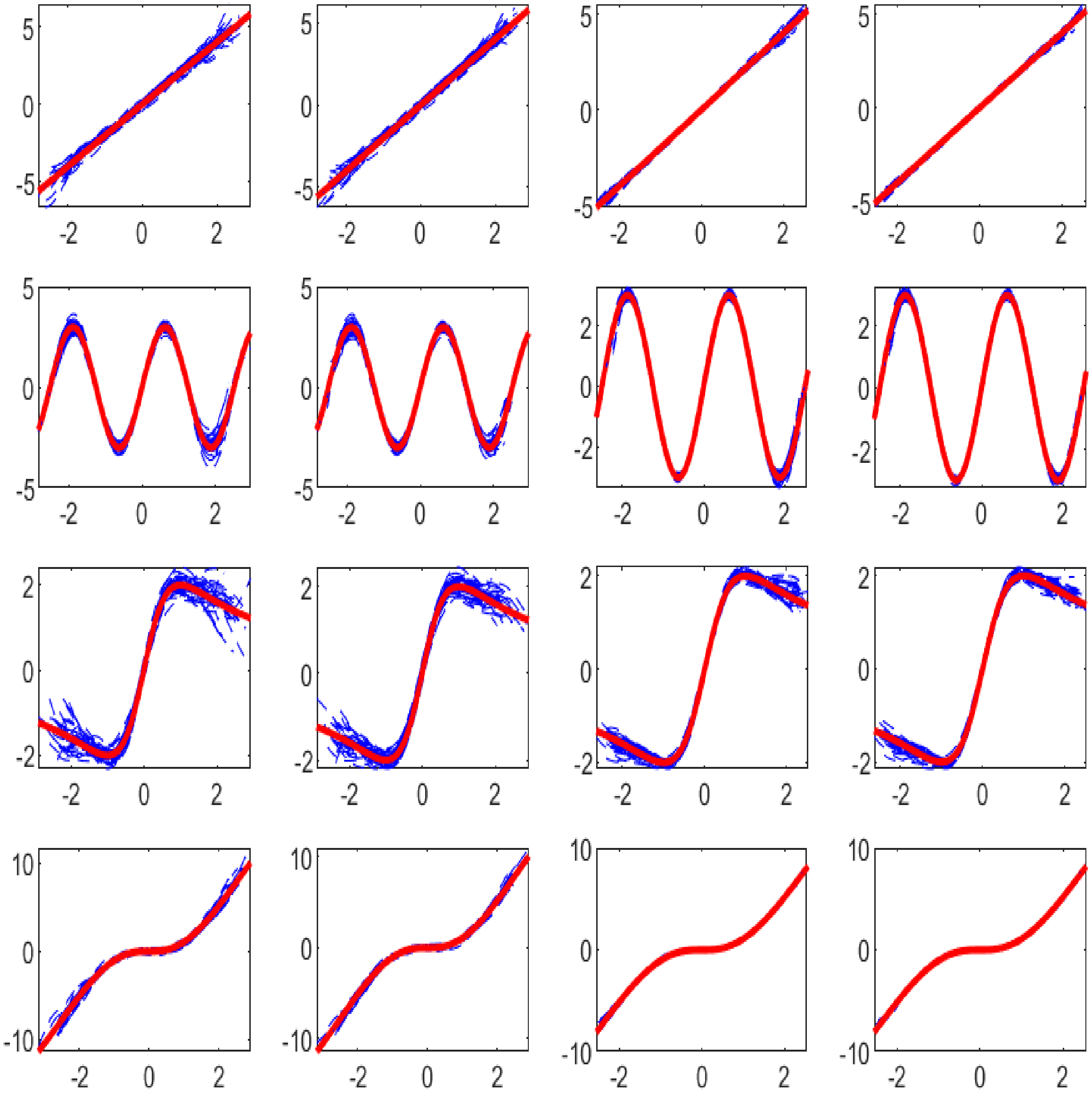

Figure 1. For lines $i = 1$ to 4, 40 estimated b_i, for $\rho = 0.25$, the R-procedure first and the oracle next, $n = 250$ for the first two columns, and $n = 1000$ for the last two columns.

where $-\|\widehat{b}_m\|_{n,R}^2$ usually stands for an estimation of the square bias term, and the penalization $\text{pen}(m)$ is proportional to the variance term. The constant κ is numerical and calibrated from preliminary simulations (see *e.g.* Comte and Genon-Catalot, 2019). For the estimator based on matrix S, called 'S-procedure', we proceed similarly

$$\widehat{\mathcal{M}}_n^{(S)} = \left\{ m \in \{1, \dots, 20\}, \sqrt{m}(\|(\widehat{\Psi}_m^{(S)})^{-1}\|_{\rm op} \vee 1) \leq c_1 \frac{n}{\log(n)} \right\}, \quad c_1 = 8 \times 256 \frac{n}{\log(n)},$$

and

$$\widehat{m}^{(S)} = \arg \min_{m \in \widehat{\mathcal{M}}_n^{(S)}} \left\{ -\|\widehat{b}_m^{(S)}\|_{n,S}^2 + \text{pen}(m) \right\}, \quad \text{pen}(m) = \kappa \frac{m}{n}.$$

The penalization constant is calibrated in all cases to $\kappa = 2.5$.

The use of the $n \times n$ matrix R^{-1} involves numerical difficulties which lead us to compute $\widehat{\Psi}_m^{(R)}$ with a formula relying of the fact that $R^{-1} = (\alpha_n - \beta_n)\text{Id}_n + \beta_n \mathbf{1}\mathbf{1}^\top$, and thus

$$\widehat{\Psi}_m^{(R)} = (\alpha_n - \beta_n) \left(\frac{1}{n} \widehat{\Phi}_m^\top \widehat{\Phi}_m \right) + n\beta_n \left(\frac{1}{n} \sum_{i=1}^n \varphi_j(X_i) \right)_{1 \leq j \leq m}^\top \left(\frac{1}{n} \sum_{i=1}^n \varphi_j(X_i) \right)_{1 \leq j \leq m}.$$

The same type of computation is done for $\widehat{\Psi}_m^{(S)}$.

The estimators are computed at points $x_1, \dots, x_K$ which are regularly spaced on an interval corresponding to the 1% and 99% quantiles of the observations, and the error (MSE) of an estimator $\widehat{b}$ of a function b is computed as the mean over 400 repetitions of the ratios

$$\frac{\sum_{k=1}^K (\widehat{b}(x_k) - b(x_k))^2}{\sum_{k=1}^K b^2(x_k)}. \tag{32}$$

In other word, we consider relative errors in order to keep a common scale for all examples. The results of the R and S data-driven procedure are compared the the performance of the so-called oracle, that is the smallest risk obtained over the models among which the selection is done; this procedure utilizes the knowledge of the true function b, this explains the term "oracle".

The results are given in Table 1, and only for X following the centered *beta*-type distribution; the result corresponding to Gaussian X_i being very similar, they are not reported. We can see that the risk decreases by a factor of order 4 when the sample size is multiplied by 4 (from $n = 250$ to $n = 1000$). The case $\rho = 0$ corresponds to the standard nonparametric least squares contrast procedure, and we can see that the order of the MSE is preserved when $\rho \neq 0$. Surprisingly, increasing ρ from 0.25 to 0.75 does not deteriorate the results but rather improves them. It is interesting to see that the method which does not require the knowledge of ρ and relies on S^{-1} works as well as the one requiring ρ: this could be expected from formula (9) which shows that $\widehat{\Psi}_m^{(R)}$ is equivalent for large n to $(1/(1-\rho))\widehat{\Psi}_m^{(S)}$, and the factor $1 - \rho$ cancels in Formula (13) giving the coefficients of the estimator for the R-procedure. Therefore, the estimator and the procedure of model selection are mainly independent of ρ. We can also note that, in mean, the selected dimensions are always smaller than the ones of the oracles. But, it is known that for model selection, choosing slightly too small dimensions is much safer than too large models; so it is not likely that this should be changed with different choices of κ or cutoff levels. The procedures and the results here are very stable.

We represent in Figure 1 the four functions and illustrate the quality of the estimators obtained by the R-procedure compared to the oracles, for $n = 250$ and $n = 1000$, in case $\rho = 0.25$. The beams of 40 estimators are in dotted-blue, and the true function in bold-red. It is worth stressing that the estimations are excellent, and there are no side-effects.

6.2. **Case $\mathbb{E}[b(X_1)]$ is unknown.** If $\mathbb{E}[b(X_1)]$ is unknown, additional obervations are required to make this term identifiable. We assume here that we also observe, for $j = 1, \dots, P$, and $i = 1, \dots, n$,

$$Y_{i,j} = b(X_i) + \sqrt{\rho}\varepsilon_{0,j} + \sqrt{1-\rho}\varepsilon_{i,j}.$$

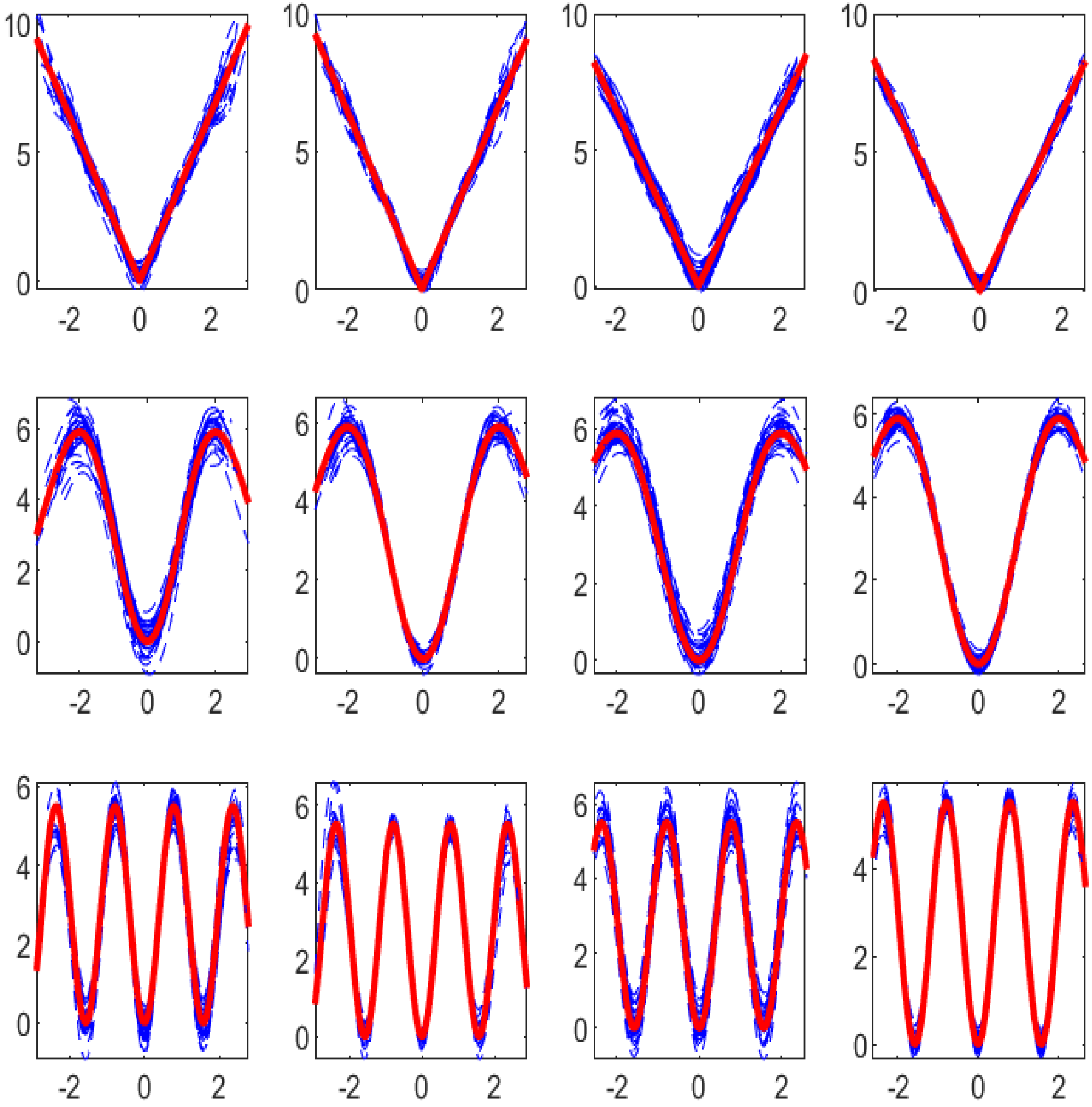

FIGURE 2. For line i, 40 estimated b_{4+i}, for $\rho = 0.5$, the R-procedure first with $P = 5$ and $P = 25$ next, $n = 250$ for the first two columns, and $n = 1000$ for the last two columns.

Then we apply the procedure to the $Y_i^\star$ defined by

$$Y_i^\star = Y_i - \bar{Y}, \quad \bar{Y} = \frac{1}{nP} \sum_{j=1}^{P} \sum_{i=1}^{n} Y_{i,j},$$

		$\rho = 0$				$\rho = 0.25$				$\rho = 0.75$			
		$n = 250$		$n = 1000$		$n = 250$		$n = 1000$		$n = 250$		$n = 1000$	
		R	O	R	O	P=5	P=25	P=5	P=25	P=5	P=25	P=5	P=25
b_5	MSE	4.29	2.95	1.17	0.93	6.19	3.74	3.57	1.53	8.69	3.16	8.14	2.06
	std	2.64	1.46	0.53	0.39	5.19	2.47	3.64	1.01	9.54	2.66	11.8	2.11
	dim	9.36	10.7	11.5	13.9	9.31	9.26	11.5	11.6	9.94	9.82	11.7	11.7
b_6	MSE	3.54	2.46	0.83	0.59	5.18	2.82	3.27	1.19	10.0	2.74	9.53	1.92
	std	4.55	0.96	0.90	0.29	5.07	1.89	3.86	0.99	12.9	2.78	12.3	2.35
	dim	10.2	11.9	12.3	14.4	5.95	5.84	7.16	7.19	6.61	6.80	7.96	7.96
b_7	MSE	7.93	5.13	2.00	1.45	9.93	6.79	5.33	2.37	18.4	5.76	14.1	3.36
	std	8.02	3.04	1.24	0.85	8.04	4.63	6.04	1.79	19.2	5.63	20.0	3.51
	dim	9.58	11.2	12.0	13.8	9.67	9.67	12.1	12.0	9.93	9.94	12.4	12.3

TABLE 2. Estimation results by R-procedure when $\mathbb{E}(b(X_1)) \neq 0$ for functions b_5, b_6, b_7, 400 repetitions, and $P = 5$ or $P = 25$ for $\rho = 0.25$ and $\rho = 0.75$. The benchmark is given for $\rho = 0$, for risk (column R) and oracle (column O). The relative MSE are computed with average of (32) and multiplied by 1000, std is also multiplied by 1000, dim is the average of selected dimensions.

and consider the relative MSE for $\widehat{b}_{\widehat{m}} + \bar{Y}$. We intend to see how large P must be to preserve the results obtained when the mean is known.

We consider the three additional functions, for Gaussian $\mathcal{N}(0,1)$ variables X_i:

$$b_5(x) = 3.25|x|, \quad b_6(x) = 4x^2 \exp(-x^2/4), \quad b_7(x) = 5.5 \sin^2(2x).$$

The results of the new experiments are given in Table 2. We first compute the relative risk as described previoulsy for one sample with $\rho = 0$, relying on standard nonparametric least-squares procedure denoted by 'R', and on the oracle denoted by 'O'. This provides a benchmark to evaluate the other cases, corresponding to $\rho = 0.25$ and $\rho = 0.75$, and R-procedure. The $Y_i^\star$ are obtained with $P = 5$ or $P = 25$. In all cases, 400 experiments are averaged. Let us mention that, with no correction, the error would explode up to orders of 100 (with the same multiplication by 1000 as in the tables), for any choice of n. We can deduce from Table 2 that the improvement is observed even for small $P = 5$, but the value of ρ has more impact here. Clearly, the case $\rho = 0.75$ is much more difficult for estimation purpose than $\rho = 0.25$, contrary to what was observed in the centered case. Also, the improvement when increasing n is attenuated, which suggests that larger P would keep improving the results. More specifically, we see that if P is too small, the estimation does not benefit from the increase of n: the MSE does not improve in the adequate factor from $n = 250$ to $n = 1000$ when $P = 5$ (see also the first and third columns in Figure 2, which are rrather similar). This means that the bias of order $\mathbb{E}(b(X_1))$ is not enough corrected when $P = 5$, which is not so surprising. Note that the selected dimensions do not depend on P. Figure 2 allows to visualize the results, for 40 repetitions and $\rho = 0.5$. Each line corresponds to a function, b_5, b_6 and b_7, and the first columns corresponds to $n = 250$ and $P = 5$, the second to the same n and $P = 25$, the last two columns are associated with $n = 1000$, again with $P = 5$ and next $P = 25$. In all these plots again, there are no side-effects at the borders, and the Hermite basis adapts to any form of function.

7. Concluding remarks

In this paper we consider the regression model (1) with correlated noises satisfying (3), *i.e.* there is a common noise to all the Y_i's. This study is inspired by models of stochastic differential equations with common noise developped in the theory of mean field games and control for models (see *e.g.* Carmona *et al.* (2016), Lacker *et al.* (2002), Maillet (2025), Delarue *et al.* (2024)) which are of the form (4) where $W_0, W_1, \dots, W_n$ are $n+1$ independent Brownian motions. Thus, W_0 is common to all equations and this creates identically distributed but correlated processes. Up to our knowledge, the regression version of these models with common noise has not been investigated up to now. It is not easy to handle and yields new results. To estimate the regression function, we proceed by projection method but we do not consider the standard least squares contrast. Instead, we use a contrast taking into account the covariance matrix of the noise vector. We study the risk of the minimum contrast estimator measured as the expectation of the empirical norm and of the theoretical norm associated with our contrast. In both cases, we obtain risks bounds where the variance term has the usual order m/n, where m is the dimension of the projection space. A rather tough difficulty occurs when proving that the empirical and the theoretical norms are equivalent for functions of one projection space. We also propose a contrast when ρ is unknown and obtain for it similar results.The theory is illustrated by numerical simulations. As the parameter $\mathbb{E}b(X_1)$ is not identifiable from the observations, we proceed first by implementing our method for functions such that $\mathbb{E}b(X_1) = 0$ for $\rho = 0$ (standard regression) and $\rho \neq 0$ for the estimator based on the contrast relying on the knowledge of ρ and for the contrast which does not require its knowledge. Then, for functions such that $\mathbb{E}b(X_1) \neq 0$, we add observations to get an estimation of $\mathbb{E}b(X_1)$ and proceed to the implementation using this estimate. The simulation results are excellent.
Note that the estimation of an unkown parameter in the drift of (4) has been studied in Genon-Catalot and Larédo (2025). The nonparametric regression model with common noise investigated here has revealed the specific difficulties of this model and it is now worth studying the nonparametric estimation of the drift in model (4) or in its discretized version.

8. Proofs

8.1. **Proof of Proposition 1.** Let us express the orthogonal projection $b_m(\mathbf{X})$ of the vector $b_A(\mathbf{X})$ on the subspace of $\mathbf{R}^n$ spanned by the vectors $\varphi_j(\mathbf{X})$, $j = 1, \dots, m$ w.r.t. the empirical scalar product $\langle . . \rangle_{n,R}$. It is given by

$$b_m(\mathbf{X}) = (\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1} b(\mathbf{X}).$$

As a consequence, by the Pythagoras theorem,

$$\begin{aligned} \|b_A - \widehat{b}_m\|_{n,R}^2 &= \|b_A - b_m\|_{n,R}^2 + \|b_m - \widehat{b}_m\|_{n,R}^2 \\ &= \inf_{t\in S_m} \|b_A - t\|_{n,R}^2 + \|(\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1}\mathbf{u}\|_{n,R}^2 \end{aligned} \tag{33}$$

Now, we have

$$\begin{aligned} \|(\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1}\mathbf{u}\|_{n,R}^2 &= \frac{1}{n}\mathbf{u}^\top R^{-1}\widehat{\Phi}_m^\top (\widehat{\Psi}_m^{(R)})^{-1} \underbrace{\widehat{\Phi}_m^\top R^{-1}\widehat{\Phi}_m}_{=n\Psi_m^{(R)}} (\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1}\mathbf{u} \\ &= \mathbf{u}^\top R^{-1}\widehat{\Phi}_m^\top (\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1}\mathbf{u}. \end{aligned}$$

This implies that

$$\begin{aligned}
\mathbb{E}(\|(\widehat{\Psi}_m^{(R)})^{-1}\widehat{\Phi}_m R^{-1}\mathbf{u}\|^2_{n,R}) &= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{u}^\top R^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top R^{-1}\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top R^{-1}\mathbf{u}\right)\right]\\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(R^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top R^{-1}\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top R^{-1}\mathbf{u}\mathbf{u}^\top\right)\right]\\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(R^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top R^{-1}\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top R^{-1}R\right)\right] = \frac{1}{n}\mathrm{Tr}(I_m)\\
&= \frac{m}{n},
\end{aligned}$$

where we used the independence of $\mathbf{u}$ and $\mathbf{X}$ and commuted the matrices inside the trace. Taking the expectation of (33) and plugging-in this result gives the announced bound. □

8.2. **Proof of Lemma 1.** The first property comes from the fact that R is a variance matrix with known eigenvalues $1-\rho$ and $1+(n-1)\rho$. Its inverse R^{-1} is diagonalizable in an orthonormal basis with eigenvalues $1/(1-\rho)$ and $1/(1+(n-1)\rho)$. The second property is obtained by taking expectation. □

8.3. **Proof of Proposition 2.** Using formula (7), we have

$$\|t\|^2_{n,R} = \alpha_n\frac{1}{n}\sum_{i=1}^n t^2(X_i) + \beta_n\frac{1}{n}\sum_{1\le i\ne k\le n} t(X_i)t(X_k).$$

As the X_i are *i.i.d.*, we get

$$\mathbb{E}(\|t\|^2_{n,R}) = \alpha_n\mathbb{E}[t^2(X_1)] + (n-1)\beta_n\{\mathbb{E}[t(X_1)]\}^2.$$

As $\beta_n < 0$, we do not stop our computation here. We further write

$$\mathbb{E}(\|t\|^2_{n,R}) = \alpha_n\mathrm{Var}(t(X_1)) + (\alpha_n + (n-1)\beta_n)\{\mathbb{E}[t(X_1)]\}^2,$$

where now $\alpha_n > 0$ and $c_n = \alpha_n + (n-1)\beta_n > 0$. Thus we have the first point of Proposition 2, from which point 2 follows straightforwardly.

We have for $i = 1,\dots,n$, $[\widehat{\Phi}_m\mathbf{a}]_i = \sum_{j=1}^m \varphi_j(X_i)a_j$. Thus

$$\begin{aligned}
\mathbf{a}^T\widehat{\Psi}_m^{(R)}\mathbf{a} &= \frac{1}{n}\left(\alpha_n\sum_{i=1}^n(\sum_{j=1}^m\varphi_j(X_i)a_j)^2 + \beta_n\sum_{i\ne i'}(\sum_{j=1}^m\varphi_j(X_i)a_j)(\sum_{k=1}^m\varphi_k(X_{i'})a_k)\right)\\
&= \frac{1}{n}\left(\alpha_n\sum_{1\le i\le n,1\le j,k\le m}\varphi_j(X_i)\varphi_k(X_i)a_ja_k + \beta_n\sum_{i\ne i',1\le j,k\le m}\varphi_j(X_i)\varphi_k(X_{i'})a_ja_k)\right)\\
&= (\alpha_n-\beta_n)\sum_{1\le j,k\le m}a_ja_k\left(\frac{1}{n}\sum_{i=1}^n\varphi_j(X_i)\varphi_k(X_i)\right)\\
&+ n\beta_n\sum_{1\le j,k\le m}a_ja_k\left(\frac{1}{n^2}\sum_{i=1}^n\varphi_j(X_i)\sum_{i=1}^n\varphi_k(X_i)\right)
\end{aligned}$$

We have $\alpha_n - \beta_n = 1/(1-\rho)$ and $n\beta_n \to -1/(1-\rho)$. Moreover, $(1/n)\sum_{i=1}^n\varphi_j(X_i)\varphi_k(X_i) \to \mathbb{E}\varphi_j(X_1)\varphi_k(X_1))$ and $(1/n^2)\sum_{i=1}^n\varphi_j(X_i)\sum_{i=1}^n\varphi_k(X_i) \to \mathbb{E}\varphi_j(X_1)\mathbb{E}\varphi_k(X_1)$. This shows that $\mathbf{a}^T\widehat{\Psi}_m^{(R)}\mathbf{a} \to \frac{1}{1-\rho}\mathbf{a}^T\Psi_m\mathbf{a}$. This ends the proof of the third point of Proposition 2.
Relation (16) is easily obtained. □

8.4. **Proof of Propositions 3 and 4.**

Proof of Proposition 3. We have (see Proposition 2 and (15)): $\Psi_m^{(R)} = \alpha_n \Psi_m + c_n \Xi_m$. This implies that

$$\alpha_n (\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m = -c_n (\Psi_m^{(R)})^{-1/2} \Xi_m (\Psi_m^{(R)})^{-1/2}.$$

Therefore

$$\|\alpha_n (\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leq |c_n| \|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \|\Xi_m\|_{\mathrm{op}}.$$

We have

$$\begin{aligned} \|\Xi_m\|_{\mathrm{op}} &= \sup_{x\in\mathbb{R}^m, \|x\|=1} x^\top \Xi_m x = \sup_{x\in\mathbb{R}^m, \|x\|=1} \left(\sum_{j=1}^m x_j \mathbb{E}[\varphi_j(X_1)] \right)^2 \\ &\leq \sup_{x\in\mathbb{R}^m, \|x\|=1} \sum_{j=1}^m x_j^2 \sum_{j=1}^m (\mathbb{E}[\varphi_j(X_1)])^2 \leq \sum_{j=1}^m (\langle \varphi_j, f\rangle])^2 \leq \|f\|^2. \end{aligned}$$

Under condition (20), we find, as $c_n \leq 1/(n\rho)$,

$$\|\alpha_n (\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leq c^\star |c_n| \|f\|^2 \frac{n}{\log(n)} \leq \frac{c^\star \|f\|^2}{\rho} \frac{1}{\log(n)}.$$

Thus, for $\log(n) \geq 2c^\star \|f\|^2/\rho$, we have

$$\|\alpha_n (\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leq \frac{1}{2}.$$

Consequetly, the eigenvalues of the matrix $\alpha_n (\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2}$ are in the interval $[1/2, 3/2]$. Thus, the eigenvalues of $[(\Psi_m^{(R)})^{-1/2} \Psi_m (\Psi_m^{(R)})^{-1/2}]^{-1}$ belong to $[2/(3\alpha_n), 2/\alpha_n]$ implying that

$$\|\Psi_m^{-1}\|_{\mathrm{op}} = \|(\Psi_m^{(R)})^{-1/2} (\Psi_m^{(R)})^{1/2} \Psi_m^{-1} (\Psi_m^{(R)})^{1/2} (\Psi_m^{(R)})^{-1/2}\|_{\mathrm{op}} \leq \frac{2}{\alpha_n} \|(\Psi_m^{(R)})^{-1/2}\|_{\mathrm{op}}^2.$$

As for $n \geq 3$, $\alpha_n \geq 1/(2(1-\rho))$, we get, for $n \geq \exp(2c^\star \|f\|^2/\rho) \vee 3$,

$$L(m) \|\Psi_m^{-1}\|_{\mathrm{op}} \leq 4(1-\rho) L(m) \|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}}^2 \leq 4c^\star (1-\rho) \frac{n}{\log(n)}.$$

This end the proof of Proposition 3. □

Proof of Proposition 4. The given expression of Ω_m is relatively standard. We have

$$\begin{aligned} \sup_{t\in S_m,\, \|t\|_R^2=1} |\|t\|_{n,R}^2 - \|t\|_R^2| &= \sup_{\mathbf{a}\in\mathbb{R}^m,\, \mathbf{a}^T \Psi_m^{(R)} \mathbf{a}=1} |\mathbf{a}^T \widehat{\Psi}_m^{(R)} \mathbf{a} - \mathbf{a}^T \Psi_m^{(R)} \mathbf{a}| \\ &= \sup_{\mathbf{b}\in\mathbb{R}^m, \mathbf{b}=(\Psi_m^{(R)})^{1/2}, \mathbf{b}^T\mathbf{b}=1} |\mathbf{b}^T (\Psi_m^{(R)})^{-1/2} \widehat{\Psi}_m^{(R)} (\Psi_m^{(R)})^{-1/2} \mathbf{b} - \mathbf{b}^T \mathbf{b}| \\ &= \|(\Psi_m^{(R)})^{-1/2} \widehat{\Psi}_m^{(R)} (\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}. \end{aligned}$$

Now, we have two steps. First, under condition (20), on Λ_m^c,

$$\begin{aligned} kc^\star \frac{n}{\log n} < L(m) \|(\widehat{\Psi}_m^{(R)})^{-1}\|_{\mathrm{op}} &\leq L(m) \|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} + L(m) \|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \\ &\leq L(m) \|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} + c^\star \frac{n}{\log n} \end{aligned}$$

Thus,

$$\Lambda_m^c \subset \{\|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \geq (k-1) \|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}}\}. \tag{34}$$

Second, for all $\alpha > 0$ and $c \in (0,1)$,

$$\begin{aligned}&\{\|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \geq \alpha \|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}}\}\\ &\subset \{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq c \wedge \alpha(1-c)\}\end{aligned}$$

Indeed,

$$\|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \leq \|(\Psi_m^{(R)})^{-1/2}\|_{\mathrm{op}}\|(\Psi_m^{(R)})^{1/2}(\widehat{\Psi}_m^{(R)})^{-1}(\Psi_m^{(R)})^{1/2} - \mathrm{Id}_m\|_{\mathrm{op}}.$$

Therefore,

$$\left\{\|(\widehat{\Psi}_m^{(R)})^{-1} - (\Psi_m^{(R)})^{-1}\|_{\mathrm{op}} \geq \alpha\|(\Psi_m^{(R)})^{-1}\|_{\mathrm{op}}\right\} \subset B := \left\{\|(\Psi_m^{(R)})^{1/2}(\widehat{\Psi}_m^{(R)})^{-1}(\Psi_m^{(R)})^{1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq \alpha\right\}$$

Now, we write $B := B_1 \cup B_2$ with

$$\begin{aligned}B_1 &= B \cap \left\{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} < c\right\}\\ B_2 &= B \cap \left\{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq c\right\}\end{aligned}$$

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

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

$$B_1 \subset \left\{\frac{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}}{1 - \|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}} \geq \alpha\right\} \cap \left\{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} < c\right\}$$

This implies $B_1 \subset \left\{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq \alpha(1-c)\right\}$.

Thus $B_1 \cup B_2 \subset \left\{\|(\Psi_m^{(R)})^{-1/2}\widehat{\Psi}_m^{(R)}(\Psi_m^{(R)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq c \wedge (1-c)\alpha\right\}$.

Now, taking into account that $\alpha = k - 1$ from (34), it remains to choose c and k such that $(k-1)(1-c) \wedge c = 3/4$. This holds for $k = 4, c = 3/4$. □

8.5. **Proof of Theorem 1.** Recall that $\|t\|_R^2 = \alpha_n \mathrm{Var}(t(X_1)) + c_n\{\mathbb{E}[t(X_1)]\}^2$. Recall that the empirical norm can be decomposed as in (16). Theorem 1 is deduced from the following lemmas.

Lemma 3. *Assume that*

$$L(m)\|\Psi_m^{-1}\|_{\mathrm{op}} \leq \mathbf{c}_p^\star(\delta)\frac{n}{\log(n)}, \quad \mathbf{c}_p^\star(\delta) \leq \mathbf{c}_p(\delta) = \frac{\delta - (1+\delta)\log(1+\delta)}{4(p+1)}, \tag{35}$$

where $\Psi_m = (\mathrm{cov}(\varphi_j(X_1), \varphi_k(X_1))_{1\leq j,k\leq m}$. Then, it holds that $\mathbb{P}((\Omega_m^{(1)}(\delta))^c) \leq 2n^{-p}$.

Lemma 4. *Recall that the basis functions are bounded by θ and take $m \leq c^\dagger n/\log(n)$. Then it holds that $\mathbb{P}((\Omega_m^{(2)})^c) \leq Cn^{-p}$, where C is a positive constant.*

We study

$$S_n(t) = \alpha_n Z_n(t) + c_n\{\mathbb{E}[t(X_1)]\}^2 + 2c_n\mathbb{E}[t(X_1)]V_n(t) \tag{36}$$

on $(\Omega_m^{(1)})^c$ (Lemma 5) and $U_n(t,t)$ on $(\Omega_m^{(2)})^c$ (Lemma 6).

Lemma 5. *Set $\Omega_m^{(1)} = \Omega_m^{(1)}(1/2)$. On $\Omega_m^{(1)}$, we have for $n \geq 192/\rho$ and for $S_n(t)$ defined by (36), that, $\forall t \in S_m$,*

$$(\frac{1}{2} - \frac{1}{8})\|t\|_R^2 \leq S_n(t) \leq (\frac{3}{2} + \frac{1}{8})\|t\|_R^2$$

Lemma 6. *Assume condition* (20)*. On* $\Omega_m^{(2)}$*,*

$$|(n-1)\beta_n U_n(t,t)| \leq \frac{1}{8}\|t\|_R^2.$$

To conclude, on $\Omega_m^{(1)} \cap \Omega_m^{(2)}$, we have proved that

$$\frac{1}{4}\|t\|_R^2 = \left(\frac{1}{2} - \frac{1}{8} - \frac{1}{8}\right)\|t\|_R^2 \leq \|t\|_{n,R}^2 \leq \left(\frac{3}{2} + \frac{1}{8} + \frac{1}{8}\right)\|t\|_R^2 = \frac{7}{4}\|t\|_R^2.$$

This means that

$$\Omega_m^{(1)} \cap \Omega_m^{(2)} \subset \Omega_m$$

and achieves the proof of Theorem 1. □

8.6. **Proof of Lemmas 3, 4, 5 and 6.**
Proof of Lemma 3. Set

$$A = \sup_{t \in S_m, Var(t(X_1))=1} |Z_n(t) - \mathrm{Var}(t(X_1))|$$

For $t = \sum_{j=1}^m a_j\varphi_j$,

$$\begin{aligned} Z_n(t) - \mathrm{Var}(t(X_1)) &= \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^m a_j(\varphi_j(X_i) - \mathbb{E}\varphi_j(X_1))\right)^2 - \mathbb{E}\left(\sum_{j=1}^m a_j(\varphi_j(X_1) - \mathbb{E}\varphi_j(X_1))\right)^2 \\ &= \mathbf{a}^T(\widehat{\psi}_m - \Psi_m)\mathbf{a} \end{aligned}$$

Therefore,

$$\begin{aligned} A &= \sup_{\mathbf{a}\in\mathbb{R}^m, \mathbf{a}^T\Psi_m\mathbf{a}=1} |\mathbf{a}^T(\widehat{\psi}_m - \Psi_m)\mathbf{a}| \\ &= \sup_{\mathbf{b}\in\mathbb{R}^m, \mathbf{b}^T\mathbf{b}=1} b^T\Psi_m^{-1/2}(\widehat{\psi}_m - \Psi_m)\Psi_m^{-1/2}\mathbf{b} = \|\Psi_m^{-1/2}\widehat{\psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \end{aligned}$$

Therefore, $\Omega_m^{(1)}(\delta)^c = \|\Psi_m^{-1/2}\widehat{\psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \delta)$. To bound $\mathbb{P}(\|\Psi_m^{-1/2}\widehat{\psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \delta)$, we use the matrix Chernov Inequality (see Tropp (2012)). For this, we define

$$K_m(X_i) = \Psi_m^{-1/2}\left((\varphi_j(X_i) - \mathbb{E}(\varphi_j(X_1)))\varphi_k(X_i) - \mathbb{E}(\varphi_k(X_1))\right)_{1\leq j,k\leq m}\Psi_m^{-1/2}.$$

As $\frac{1}{n}\sum_{i=1}^n \mathbb{E}K_m(X_i) = \mathrm{Id}_m$, the Tropp inequalities write

$$\begin{aligned} \mathbb{P}(\lambda_{min}\left(\frac{1}{n}\sum_{i=1}^n K_m(X_i)\right) \leq 1-\delta) &\leq m\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{1/r} \\ \mathbb{P}(\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^n K_m(X_i)\right) \geq 1+\delta) &\leq m\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{1/r} \end{aligned}$$

where λ_{min} (resp. λ_{max}) denotes the smallest (resp largest) eigenvalue of the matrix and r is an upper bound for the largest eigenvalue of $\frac{1}{n}K_m(X_i)$. This implies

$$\mathbb{P}(\Omega_m^{(1)}(\delta)^c) \leq 2m\exp\left(-c(\delta)/r\right)$$

where $c(\delta)=(1+\delta)\log(1+\delta)-\delta$.
We must now compute the bound r. For this, we look at

$$\begin{aligned}
\sup_{\mathbf{a}^T\mathbf{a}=1}\frac{1}{n}\mathbf{a}^TK_m(X_i)\mathbf{a} &= \frac{1}{n}\sup_{\mathbf{b}=\psi_m^{-1/2}\mathbf{a},\mathbf{a}^T\mathbf{a}=1}\mathbf{b}^T\left((\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_1)))\varphi_k(X_i)-\mathbb{E}(\varphi_k(X_1))\right)_{1\le j,k\le m}\mathbf{b}\\
&= \frac{1}{n}\sup_{\mathbf{b}=\psi_m^{-1/2}\mathbf{a},\mathbf{a}^T\mathbf{a}=1}\left(\sum_{j=1}^m b_j(\varphi_j(X_i)-\mathbb{E}\varphi_j(X_1))\right)^2\\
&\le \frac{1}{n}\|\mathbf{b}\|_m^2\sum_j(\varphi_j(X_i)-\mathbb{E}\varphi_j(X_1))^2\le\frac{1}{n}\|\Psi_m\|_{\text{op}}^{-1}4L(m)
\end{aligned}$$

Therefore, under the condition $L(m)\|\Psi_m\|_{\text{op}}^{-1}\le c_p^\star(\delta)(n/\log(n))$ with $c_p^\star(\delta)\le\frac{c(\delta)}{4(p+1)}$, we obtain $\mathbb{P}(\Omega_m^{(1)})(\delta)^c)\le 2n^{-p}$. $\square$

Proof of Lemma 4. We proceed in two steps, first the application of the deviation for U-statistics to $U_n(\varphi_j,\varphi_k)$ and then the use of the result to handle $U_n(t)=U_n(t,t)$.

Step 1 We intend to apply Theorem 3.4 in Houdré and Reynaud-Bouret (2003), and prove that, for $\delta_{j,k}$ given by (40) below with $q=p+2$,

$$\mathbb{P}\left(|U_n(\varphi_j,\varphi_k)|\ge\delta_{j,k}\right)\le\frac{1}{n^{p+2}}.$$

We set,

$$\begin{aligned}
g(X_i,X_\ell) &= g_{i,\ell}(X_i,X_\ell)=\left[\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_1))\right]\left[\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_1))\right]\\
&\qquad+\left[\varphi_j(X_\ell)-\mathbb{E}(\varphi_j(X_1))\right]\left[\varphi_k(X_i)-\mathbb{E}(\varphi_k(X_1))\right].
\end{aligned}$$

It satisfies

$$\mathbb{E}[g(X_i,X_\ell)|X_i]=\mathbb{E}[g(X_i,X_\ell)|X_\ell]=0$$

and $U_n(\varphi_j,\varphi_k)=\sum_{i=2}^n\sum_{\ell=1}^{i-1}g(X_i,X_\ell)$.

To apply Theorem 3.4 in Houdré and Reynaud-Bouret (2003) to $U_n(\varphi_j,\varphi_k)$, we compute the terms denoted by A,B,C,D in this paper. First, we find $\|g\|_\infty\le 8\|\varphi_j\varphi_k\|_\infty\le 8\theta^2:=A$.

Next B^2 is a bound on

$$\max\left\{\sup_{i,x}\sum_{\ell=1}^{i-1}[\mathbb{E}^{(\ell)}(g(x,X_\ell))]^2,\ \sup_{\ell,x}\sum_{i=\ell+1}^{n}[\mathbb{E}_{(\ell)}(g(X_i,x))]^2\right\},\tag{37}$$

where $[\mathbb{E}^{(\ell)}(g(X_i,X_\ell))=\mathbb{E}[g(X_i,X_\ell)|X_i]$ and $[\mathbb{E}_{(i)}(g(X_i,X_\ell))=\mathbb{E}[g(X_i,X_\ell)|X_\ell]$. We find that (37) is less than

$$\begin{aligned}
4(n-1)(\|\varphi_j\|_\infty^2\text{Var}(\varphi_k(X_1))+\|\varphi_k\|_\infty^2\text{Var}(\varphi_j(X_1))) &\le 4(n-1)(\|\varphi_k\|_\infty^2\int\varphi_j^2f+\|\varphi_j\|_\infty^2\int\varphi_k^2f)\\
&\le 4(n-1)\theta^2(\int\varphi_j^2f+\int\varphi_k^2f):=B^2.
\end{aligned}$$

The third quantity C^2 is defined as a bound on

$$\sum_{i=2}^n\sum_{\ell=1}^{i-1}\mathbb{E}[g^2(X_i,X_\ell)]=n(n-1)\text{Var}(\varphi_j(X_1))\text{Var}(\varphi_k(X_1))\le n(n-1)\int\varphi_j^2f\int\varphi_k^2f.$$

Therefore, we take

$$C^2 := n(n-1)\int \varphi_j^2 f \int \varphi_k^2 f.$$

Lastly the most difficult part is to find D which is defined as an upper bound on

$$\sup\left\{\mathbb{E}\left(\sum_{i=2}^{n}\sum_{\ell=1}^{i-1} a_i(X_i)b_\ell(X_\ell)g(X_i,X_\ell)\right),\ \mathbb{E}\left(\sum_{i=2}^{n} a_i^2(X_i)\right)\leq 1, \mathbb{E}\left(\sum_{\ell=1}^{n-1} b_\ell^2(X_\ell)\right)\leq 1\right\}.$$

By the proof below, we obtain

$$D := 2\sqrt{n(n-1)}\sqrt{\int \varphi_j^2 f \int \varphi_k^2 f}. \tag{38}$$

Proof of the bound for D.
We consider the first half of g. Assume that the functions $a_i(x), b_\ell(x)$ are such that $\mathbb{E}\left[\sum_{i=2}^n a_i^2(X_i)\right]\leq 1$ and $\mathbb{E}\left[\sum_{\ell=1}^{n-1} b_\ell^2(X_\ell)\right]\leq 1$. We have

$$\mathbb{E}\left[\sum_{i=2}^{n}\sum_{\ell=1}^{i-1} a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))\right]$$

$$= \sum_{i=2}^{n}\sum_{\ell=1}^{i-1}\mathbb{E}[a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))]\mathbb{E}[b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))]$$

$$= \mathbb{E}\left[\sum_{i=2}^{n} a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))\mathbb{E}[\sum_{\ell=1}^{i-1} b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))]\right]. \tag{39}$$

Then

$$\begin{aligned}\left|\mathbb{E}[\sum_{\ell=1}^{i-1} b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))\right| &\leq \mathbb{E}\left[\left(\sum_{\ell=1}^{i-1} b_\ell^2 X_\ell)\sum_{\ell=1}^{i-1}(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))^2\right)^{1/2}\right]\\ &\leq \left[\mathbb{E}\left(\sum_{\ell=1}^{i-1} b_\ell^2(X_\ell)\right)\mathbb{E}\left(\sum_{\ell=1}^{i-1}(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))^2\right)\right]^{1/2}\end{aligned}$$

using that $\mathbb{E}(\sqrt{UV})\leq\sqrt{\mathbb{E}(U)\mathbb{E}(V)}$ for two random variables $U,V\geq 0$. So, we have obtained

$$\left|\mathbb{E}[\sum_{\ell=1}^{i-1} b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))\right|\leq\sqrt{(i-1)\mathrm{Var}(\varphi_k(X_1))}\leq\sqrt{(i-1)\int\varphi_k^2 f}$$

Plugging this in (39), we get

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=2}^{n}\sum_{\ell=1}^{i-1} a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))\right]\\ \leq\ &\sqrt{(n-1)\int\varphi_k^2 f}\ \mathbb{E}\left[\sum_{i=2}^{n}|a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))|\right].\end{aligned}$$

Dealing with the above term containing the a_i's as with the term containing the b_ℓ's yields

$$\mathbb{E}\left[\sum_{i=2}^{n}\sum_{\ell=1}^{i-1} a_i(X_i)(\varphi_j(X_i)-\mathbb{E}(\varphi_j(X_i))b_\ell(X_\ell)(\varphi_k(X_\ell)-\mathbb{E}(\varphi_k(X_\ell))\right]\leq\sqrt{n(n-1)}\sqrt{\int\varphi_k^2 f\int\varphi_j^2 f}.$$

This gives the announced bound on D.

Therefore, choosing $u = q\log(n)$ and $\varepsilon = 4$ in Theorem 3.4 of Houdré and Reynaud-Bouret (2003), gives that with probability larger than $1 - 2 \times 2.77n^{-q}$,

$$\frac{U_n(\varphi_j,\varphi_k)}{n(n-1)} \leq \frac{c_0}{n(n-1)}\left(C\sqrt{q\log(n)} + D(q\log(n)) + B(q\log(n))^{3/2} + A(q\log(n))^2\right),$$

where c_0 is a numerical constant, $c_0 = 62.3$ for $\varepsilon = 4$ (given by the worst constant $\beta(\epsilon)$ for $\epsilon = 4$ in the quoted paper). That is, plugging the values of A, B, C, D given above and using that $\|\varphi_j\|_\infty \leq \theta$,

$$\begin{aligned}\frac{U_n(\varphi_j,\varphi_k)}{n(n-1)} \quad \leq \quad & 8c_0\, q^2 \left(\frac{\sqrt{\int\varphi_j^2 f\int\varphi_k^2 f}}{\sqrt{n(n-1)}}\sqrt{\log(n)} + \frac{\sqrt{\int\varphi_j^2 f\int\varphi_k^2 f}}{\sqrt{n(n-1)}}\log(n)\right.\\ & \left. + \frac{\theta\left(\sqrt{\int\varphi_k^2 f} + \sqrt{\int\varphi_j^2 f}\right)}{n\sqrt{n-1}}\log^{3/2}(n) + \frac{\theta^2}{n(n-1)}\log^2(n)\right) := \delta_{j,k}\end{aligned} \tag{40}$$

Step 2. Then, using that for $t = \sum_{j=1}^m a_j\varphi_j$,

$$U_n(t,t) = \sum_{1\leq j,k\leq m} a_j a_k U_n(\varphi_j,\varphi_k) \leq \|t\|^2\left(\sum_{1\leq j,k\leq m} U_n^2(\varphi_j,\varphi_k)\right)^{1/2},$$

we have

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in S_m, t\neq 0}\left|\frac{U_n(t,t)}{\|t\|^2}\right| \geq \sqrt{\sum_{1\leq j,k\leq m}\delta_{j,k}^2}\right) \quad & \leq \quad \mathbb{P}\left(\sum_{1\leq j,k\leq m} U_n^2(\varphi_j,\varphi_k) \geq \sum_{1\leq j,k\leq m}\delta_{j,k}^2\right)\\ & \leq \quad \mathbb{P}\left(\exists j,k, \text{ such that } U_n^2(\varphi_j,\varphi_k) \geq \delta_{j,k}^2\right)\\ & \leq \quad \sum_{1\leq j,k\leq m}\mathbb{P}\left(U_n^2(\varphi_j,\varphi_k) \geq \delta_{j,k}^2\right)\\ & \leq \quad \sum_{1\leq j,k\leq m}\mathbb{P}\left(|U_n(\varphi_j,\varphi_k)| \geq \delta_{j,k}\right) \leq n^{-p}.\end{aligned}$$

Now as the basis is bounded by θ, $q = p+2$, and $\delta_{j,k}$ defined by (40), we get that

$$\begin{aligned}\sum_{1\leq j,k\leq m}\delta_{j,k}^2 \leq 4(8c_0)^2\,(p+2)^4 & \left(2L^2(m)\frac{\log^2(n)}{n^2} + 4\theta^2 mL(m)\frac{\log^3(n)}{n^3} + \theta^4\frac{m^2\log^4(n)}{n^4}\right)\\ \leq \quad 4(8c_0)^2(p+2)^4 & \left(4L^2(m)\frac{\log^2(n)}{n(n-1)} + 3m^2\theta^4\frac{\log^4(n)}{n^2(n-1)^2}\right),\end{aligned}$$

where we use that

$$2\theta^2 mL(m)\frac{\log^3(n)}{n^3} = 2\frac{L(m)\log(n)}{\sqrt{n(n-1)}} \times \frac{m\theta^2\log^2(n)}{n\sqrt{n(n-1)}} \leq \frac{L^2(m)\log^2(n)}{n(n-1)} + \frac{m^2\theta^4\log^4(n)}{n^2(n-1)^2}.$$

For $n \geq 5$, we have $1/(n-1) \leq (5/4)(1/n)$ and thus

$$\sum_{1\leq j,k\leq m}\delta_{j,k}^2 \leq 4(8c_0)^2(p+2)^4\left(5L^2(m)\frac{\log^2(n)}{n^2} + 3\,(\frac{5}{4})^2\, m^2\theta^4\frac{\log^4(n)}{n^4}\right).$$

Therefore, for $m \leq c^\dagger n/\log(n)$, we get

$$\sqrt{\sum_{1\leq j,k\leq m} \delta_{j,k}^2} \leq 24c_0(\theta^2 \vee 1)(p+2)^2(L(m)+c^\dagger)\frac{\log(n)}{n}.$$

We obtain

$$\mathbb{P}\left(\sup_{t\in S_m, t\neq 0}\left|\frac{U_n(t,t)}{\|t\|^2}\right| \geq 24c_0(\theta^2\vee 1)\,(p+2)^2(L(m)+c^\dagger)\frac{\log(n)}{n}\right) \leq cn^{-p},$$

and this ends the proof of Lemma 4. □

Proof of Lemma 5. On $\Omega_m^{(1)}$, we have $\forall t \in S_m$,

$$\frac{1}{2}\mathrm{Var}(t(X_1)) \leq Z_n(t) \leq \frac{3}{2}\mathrm{Var}(t(X_1)).$$

Therefore,

$$\alpha_n\frac{1}{2}\mathrm{Var}(t(X_1)) + \frac{1}{2}c_n(\mathbb{E}t(X_1))^2 \quad\leq\quad \alpha_n Z_n(t) + c_n(\mathbb{E}t(X_1))^2 \leq \alpha_n\frac{3}{2}\mathrm{Var}(t(X_1)) + \frac{3}{2}c_n(\mathbb{E}t(X_1))^2$$

We thus have

$$\frac{1}{2}\|t\|_R^2 \leq \alpha_n Z_n(t) + c_n(\mathbb{E}t(X_1))^2 \leq \frac{3}{2}\|t\|_R^2. \tag{41}$$

Now, note that

$$c_n = \frac{1}{1-\rho+n\rho} \leq \frac{1}{n\rho} \text{ and } \quad \forall n\geq 3,\ \alpha_n = \frac{1}{1-\rho}\,\frac{1+(n-2)\rho}{1+(n-1)\rho} \geq \frac{1}{2}\frac{1}{(1-\rho)}.$$

Then using that $2|xy| \leq (1/\tau)x^2 + \tau y^2$ for all $\tau > 0$, and $V_n^2(t) \leq Z_n(t)$, we get that, $\forall \tau > 0$,

$$\begin{aligned}
2c_n|\mathbb{E}[t(X_1)]V_n(t)| &= c_n^{1/2}\,2[c_n^{1/2}\mathbb{E}[t(X_1)]]\,V_n(t) \leq c_n^{1/2}\left(\frac{1}{\tau}c_n(\mathbb{E}[t(X_1)])^2 + \tau V_n^2(t)\right)\\
&\leq \frac{1}{\sqrt{n\rho}}\left(\frac{1}{\tau}c_n(\mathbb{E}[t(X_1)])^2 + \tau Z_n(t)\right)\\
&\leq \frac{1}{\sqrt{n\rho}}\left(\frac{1}{\tau}c_n(\mathbb{E}[t(X_1)])^2 + \frac{3}{2}\tau\mathrm{Var}(t(X_1))\right)\\
&\leq \frac{1}{\sqrt{n\rho}}\left(\frac{1}{\tau}c_n(\mathbb{E}[t(X_1)])^2 + 3(1-\rho)\tau\alpha_n\mathrm{Var}(t(X_1))\right)\\
&\leq \frac{1}{\sqrt{n\rho}}\left(\frac{1}{\tau}c_n(\mathbb{E}[t(X_1)])^2 + 3\tau\alpha_n\mathrm{Var}(t(X_1))\right).
\end{aligned}$$

Choosing $\tau = 1/\sqrt{3}$, we obtain that on $\Omega_m^{(1)}$

$$2c_n|\mathbb{E}[t(X_1)]V_n(t)| \leq \frac{\sqrt{3}}{\sqrt{n\rho}}\left(c_n(\mathbb{E}[t(X_1)])^2 + \alpha_n\mathrm{Var}(t(X_1))\right) = \frac{\sqrt{3}}{\sqrt{n\rho}}\|t\|_R^2. \tag{42}$$

We choose n such that $\frac{\sqrt{3}}{\sqrt{n\rho}} \leq 1/8$, *i.e.*, $n \geq 192/\rho$ and joining (41) and (42), we get the result of Lemma 5. □

Proof of Lemma 6. Note that for $t \in S_m$, $\|t\| \neq 0$ and $\mathrm{Var}[t(X_1)] \neq 0$,

$$\sup_{t\in S_m}\frac{|U_n(t,t)|}{\mathrm{Var}(t(X_1))} \leq \sup_{t\in S_m}\frac{|U_n(t,t)|}{\|t\|^2}\sup_{t\in S_m}\frac{\|t\|^2}{\mathrm{Var}(t(X_1))} \leq \|\Psi_m^{-1}\|_{\mathrm{op}}\sup_{t\in S_m}\frac{|U_n(t,t)|}{\|t\|^2}.$$

On $\Omega_m^{(2)}$, we use $L(m)+c^\dagger \leq 2(c^\dagger \vee 1)L(m)$ as $L(m) \geq 1$, so that, setting $\mathbf{c}_1 = 2 \times 24 c_0(\theta^2 \vee 1)\,(p+2)^2(c^\dagger \vee 1)$, it follows that

$$|U_n(t,t)| \leq \mathbf{c}_1 \|\Psi_m^{-1}\|_{\mathrm{op}} L(m) \frac{\log(n)}{n} \mathrm{Var}(t(X_1)).$$

So under condition (35), we have that, on $\Omega_m^{(2)}$,

$$|U_n(t,t)| \leq \mathbf{c}_1 \mathbf{c}_p^\star(1/2) \mathrm{Var}(t(X_1)).$$

We note that

$$|\frac{(n-1)\beta_n}{\alpha_n}| = \frac{(n-1)\rho}{1+(n-2)\rho} = \frac{(n-1)\rho}{1-\rho+(n-1)\rho} \leq 1,$$

and thus on $\Omega_m^{(2)}$,

$$|(n-1)\beta_n U_n(t,t)| \leq \mathbf{c}_1 \mathbf{c}_p^\star(1/2)\alpha_n \mathrm{Var}(t(X_1)) \leq \mathbf{c}_1 \mathbf{c}_p^\star(1/2)\|t\|_R^2.$$

We choose $\mathbf{c}_p^\star(1/2)$ defined by (35) as

$$\mathbf{c}_p^\star(1/2) = \mathbf{c}_p(1/2) \wedge (1/16\mathbf{c}_1).$$

By Proposition 3, the stability condition (20) implies $L(m)\|\Psi_m^{-1}\|_{\mathrm{op}} \leq 4c^\star(1-\rho)\frac{n}{\log(n)}$. So we choose

$$c^\star = \frac{1}{4(1-\rho)}\left(\mathbf{c}_p(1/2) \wedge \frac{1}{16\mathbf{c}_1}\right),$$

and get that on $\Omega_m^{(2)}$, $|(n-1)\beta_n U_n(t,t)| \leq \frac{1}{8}\|t\|_R^2$ as soon as

$$2\mathbf{c}_1\mathbf{c}_p^\star(1/2) \leq \frac{1}{8}. \quad \square$$

8.7. **Proof of Theorem 2.** Using Proposition 4 and Theorem 1:

$$\mathbb{E}[\|\widetilde{b}_m - b\|_R^2] = \mathbb{E}[\|\widehat{b}_m - b\|_R^2 \mathbf{1}_{\Lambda_m}] + \|b\|_R^2 \mathbb{P}(\Lambda_m^c) \leq T_1 + \frac{\|b\|_R^2}{n^p},$$

where

$$T_1 := \mathbb{E}[\|\widehat{b}_m - b\|_R^2 \mathbf{1}_{\Lambda_m}].$$

For any $t \in S_m$,

$$T_1 \leq 2T_{1,1} + 2\|t-b\|_R^2,$$

where $T_{1,1} = \mathbb{E}[\|\widehat{b}_m - t\|_R^2 \mathbf{1}_{\Lambda_m}]$. On Ω_m, for all functions $t \in S_m$,

$$\frac{1}{4}\|t\|_R^2 \leq \|t\|_{n,R}^2 \leq \frac{7}{4}\|t\|_R^2.$$

Therefore, we can write

$$\begin{aligned}
T_{1,1} &= \mathbb{E}[\|\widehat{b}_m - t\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m}] + \mathbb{E}[\|\widehat{b}_m - t\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] \\
&\leq 4\mathbb{E}[\|\widehat{b}_m - t\|_{n,R}^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m}] + \mathbb{E}[\|\widehat{b}_m - t\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] \\
&\leq 8\mathbb{E}[\|\widehat{b}_m - b\|_{n,R}^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m}] + 8\mathbb{E}[\|b-t\|_{n,R}^2] + \mathbb{E}[\|\widehat{b}_m - t\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}]
\end{aligned}$$

Now we use Proposition 1, choose $t = \Pi_{S_m,R}b$ as the minimizer of $\|b-t\|_R^2$ in S_m and obtain

$$\mathbb{E}[\|\widetilde{b}_m - b\|_R^2] \leq 18 \inf_{t \in S_m} \|t-b\|_R^2 + 16\frac{m}{n} + \frac{\|b\|_R^2}{n^p} + \mathbb{E}[\|\widehat{b}_m - \Pi_{S_m,R}b\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}]. \tag{43}$$

Let us show that the last term is negligible.

$$\mathbb{E}[\|\widehat{b}_m - \Pi_{S_m,R}b\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] \leq 2\mathbb{E}[\|\widehat{b}_m\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] + 2\|b\|_R^2 \mathbb{P}(\Omega_m^c).$$

The last term is negligible: $2\|b\|_R^2\mathbb{P}(\Omega_m^c) \lesssim n^{-p}$. First note that (see (12)) $(\widehat{\Phi}_m^T R^{-1}\widehat{\Phi}_m)^{-1} = \frac{1}{n}(\widehat{\Psi}_m^{(R)})^{-1}$ and recall (13). On Λ_m,

$$\begin{aligned}
\|\widehat{b}_m\|_R^2 &= \widehat{\mathbf{a}}_m^\top \Psi_m^{(R)} \widehat{\mathbf{a}}_m = \frac{1}{n^2} Y^\top R^{-1}\widehat{\Phi}_m \left(\widehat{\Psi}_m^{(R)}\right)^{-1} \Psi_m^{(R)} \left(\widehat{\Psi}_m^{(R)}\right)^{-1} \widehat{\Phi}_m^T R^{-1} Y \\
&\leq \frac{1}{n^2}\| \left(\widehat{\Psi}_m^{(R)}\right)^{-1} \|_{\mathrm{op}}^2 \|\Psi_m^{(R)}\|_{\mathrm{op}} \|\widehat{\Phi}_m^T R^{-1} Y\|_2^2 \\
&\leq \frac{1}{n^2}\| \left(\widehat{\Psi}_m^{(R)}\right)^{-1} \|_{\mathrm{op}}^2 \|\Psi_m^{(R)}\|_{\mathrm{op}} \|R^{-1}\|_{\mathrm{op}}^2 \|\widehat{\Phi}_m^T \widehat{\Phi}_m\|_{\mathrm{op}} \|Y\|^2 \\
&\leq \frac{1}{n^2}\left(c^\star \frac{n}{L(m)\log(n)}\right)^2 \|\Psi_m^{(R)}\|_{\mathrm{op}} \left(\frac{1}{1-\rho}\right)^2 \|\widehat{\Phi}_m^T \widehat{\Phi}_m\|_{\mathrm{op}} \|Y\|^2.
\end{aligned}$$

We have

$$\begin{aligned}
\|\widehat{\Phi}_m^T \widehat{\Phi}_m\|_{\mathrm{op}} &= \sup_{\mathbf{x}\in\mathbb{R}^m, \|x\|=1} \left[\sum_{i=1}^n \left(\sum_{j=1}^m x_j \varphi_j(X_i)\right)^2\right] \\
&\leq \sup_{\mathbf{x}\in\mathbb{R}^m, \|\mathbf{x}\|=1} \left[\sum_{i=1}^n \sum_{j=1}^m x_j^2 \sum_{j=1}^m \varphi_j^2(X_i)\right] \leq nL(m).
\end{aligned}$$

Moreover

$$\|\Psi_m^{(R)}\|_{\mathrm{op}} = \frac{1}{n}\sup_{\mathbf{x}\in\mathbb{R}^m, \|\mathbf{x}\|=1} \mathbb{E}\left[\mathbf{x}^T \widehat{\Phi}_m^T R^{-1}\widehat{\Phi}_m \mathbf{x}\right] \leq \frac{1}{n}\|R^{-1}\|_{\mathrm{op}} \|\widehat{\Phi}_m^T \widehat{\Phi}_m\|_{\mathrm{op}} \leq \frac{L(m)}{1-\rho}.$$

Therefore, on Λ_m,

$$\|\widehat{b}_m\|_R^2 \leq \frac{(c^\star)^2}{(1-\rho)^3}\frac{n}{(\log(n))^2}\|Y\|^2.$$

Next, we have

$$\mathbb{E}[\|\widehat{b}_m\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] \leq \frac{(c^\star)^2}{(1-\rho)^3}\frac{n}{(\log(n))^2}\mathbb{E}^{1/2}(\|Y\|^4)\sqrt{\mathbb{P}(\Omega_m^c)}.$$

As $\mathbb{E}(\|Y\|^4) = \mathbb{E}\left[\left(\sum_{i=1}^n Y_i^2\right)^2\right] \leq n\mathbb{E}\left[\sum_{i=1}^n Y_i^4\right] = n^2\mathbb{E}[Y_1^4]$, we obtain

$$\mathbb{E}[\|\widehat{b}_m\|_R^2 \mathbf{1}_{\Lambda_m}\mathbf{1}_{\Omega_m^c}] \leq \frac{(c^\star)^2}{(1-\rho)^3}\frac{n^2}{(\log(n))^2}\mathbb{E}^{1/2}(Y_1^4)\sqrt{\mathbb{P}(\Omega_m^c)}.$$

This term is therefore of order $n^{-(p/2)+2}/\log^2(n)$, that is less than $1/n$ for $p \geq 6$. Plugging this in (43), we obtain that, for $p \geq 6$,

$$\mathbb{E}[\|\widetilde{b}_m - b\|_R^2] \leq 18 \inf_{t\in S_m} \|t-b\|_R^2 + 16\frac{m}{n} + \frac{C}{n}$$

for some positive constant C. This gives the result of Theorem 2. □

8.8. **Proofs of the results of Section 4.**

Proof of Proposition 5. The line is the same as the proof of Proposition 1. Let us express the orthogonal projection $b_m^{(S)}(\mathbf{X})$ of the vector $b_A(\mathbf{X})$ on the subspace of $\mathbf{R}^n$ spanned by the vectors $\varphi_j(\mathbf{X})$, $j = 1, \dots, m$ w.r.t. the empirical scalar product $\langle . . \rangle_{n,S}$. It is given by

$$b_m^{(S)}(\mathbf{X}) = (\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1} b(\mathbf{X}).$$

As a consequence, by the Pythagoras theorem,

$$\begin{aligned}\|b_A - \widehat{b}_{m,S}\|_{n,S}^2 &= \|b_A - b_m^{(S)}\|_{n,S}^2 + \|b_m^{(S)} - \widehat{b}_{m,S}\|_{n,S}^2 \\ &= \inf_{t\in S_m} \|b_A - t\|_{n,S}^2 + \|(\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1}\mathbf{u}\|_{n,S}^2\end{aligned}$$

Now, we have

$$\begin{aligned}\|(\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1}\mathbf{u}\|_{n,S}^2 &= \frac{1}{n}\mathbf{u}^\top S^{-1}\widehat{\Phi}_m^\top(\widehat{\Psi}_m^{(S)})^{-1}\underbrace{\widehat{\Phi}_m^\top S^{-1}\widehat{\Phi}_m}_{=n\Psi_m^{(S)}}(\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1}\mathbf{u} \\ &= \mathbf{u}^\top S^{-1}\widehat{\Phi}_m^\top(\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1}\mathbf{u}.\end{aligned}$$

This implies that

$$\begin{aligned}\mathbb{V}_m := \mathbb{E}(\|(\widehat{\Psi}_m^{(S)})^{-1}\widehat{\Phi}_m S^{-1}\mathbf{u}\|_{n,S}^2) &= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{u}^\top S^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top S^{-1}\,\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top S^{-1}\mathbf{u}\right)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(S^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top S^{-1}\,\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top S^{-1}\mathbf{u}\mathbf{u}^\top\right)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(S^{-1}\widehat{\Phi}_m\left(\widehat{\Phi}_m^\top S^{-1}\,\widehat{\Phi}_m\right)^{-1}\widehat{\Phi}_m^\top S^{-1}R\right)\right].\end{aligned}$$

We write it as $\mathbb{V}_m = \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left(R^{1/2}S^{-1}\widehat{\Phi}_m(\widehat{\Phi}_m^\top S\widehat{\Phi}_m)^{-1}\widehat{\Phi}_m^\top S^{-1}R^{1/2}\right)\right]$, where $R^{1/2}$ is a symmetric square root of R. Now, R and S^{-1} are diagonalisable in the same orthonormal basis, meaning that there exist $P \in \mathcal{M}_n(\mathbb{R})$ such that $P^\top P = PP^\top = I_n$ and $R = PD_1P^\top$, $S^{-1} = PD_2P^\top$ with $D_1 = \mathrm{diag}((1-\rho),\ldots,(1-\rho),1+(n-1)\rho)$ and $D_2 = \mathrm{diag}(1+1/n,\ldots,1+1/n,1/n)$ (See section 9). As as consequence

$$R^{1/2}S^{-1} = P\Delta P^\top, \quad \Delta = \mathrm{diag}\left(\sqrt{1-\rho}(1+\frac{1}{n}),\ldots,\sqrt{1-\rho}(1+\frac{1}{n}),\frac{\sqrt{1+(n-1)\rho}}{n}\right)$$

Now we note that for M a symmetric nonnegative matrix

$$\mathrm{Tr}(P\Delta P^\top MP\Delta P^\top) = \mathrm{Tr}(\Delta P^\top MP\Delta) = \sum_{i=1}^n \Delta_i^2\left[PMP^\top\right]_{i,i}.$$

It is easy to see that $PMP^\top$ is also symmetric and nonnegative and thus $\left[PMP^\top\right]_{i,i} \geq 0$. As $\Delta_i \leq \sqrt{2[D_2]_i}$ using that

$$\sqrt{(1-\rho)(1+\frac{1}{n})} \leq \sqrt{2}, \quad \sqrt{\frac{1+(n-1)\rho}{n}} = \sqrt{\rho+\frac{1-\rho}{n}} \leq 1,$$

we obtain $\mathrm{Tr}(P\Delta P^\top MP\Delta P^\top) \leq 2\mathrm{Tr}(PD_2^{1/2}P^\top MPD_2^{1/2}P^\top)$. As a consequence,

$$\mathrm{Tr}\left(R^{1/2}S^{-1}\widehat{\Phi}_m(\widehat{\Phi}_m^\top S^{-1}\widehat{\Phi}_m)^{-1}\widehat{\Phi}_m^\top S^{-1}R^{1/2}\right) \leq 2\mathrm{Tr}\left(S^{-1/2}\widehat{\Phi}_m(\widehat{\Phi}_m^\top S^{-1}\widehat{\Phi}_m)^{-1}\widehat{\Phi}_m^\top S^{-1/2}\right) = 2m.$$

It follows that $\mathbb{V}_m \leq 2m/n$ which gives the result.□

Proof of Theorem 3.
To prove Theorem 3, we follow the same steps of in the proof of Theorem 2. Let us define

$$\Omega_{m,S} = \left\{\forall t \in S_m, t \neq 0, \left|\frac{\|t\|_{n,S}^2}{\|t\|_S^2} - 1\right| \le \frac{3}{4}\right\}. \tag{44}$$

As in Proposition 4, it holds that

$$\Omega_{m,S} = \left\{\|((\Psi_m^{(S)})^{-1/2}\widehat{\Psi}_m^{(S)}(\Psi_m^{(S)})^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \le \frac{3}{4}\right\}$$

and under condition (28), we have: $\Lambda_{m,S}^c \subset \Omega_{m,S}^c$. Following the line of the proof of Theorem 2, we get that the first part holds, as we can bound $\|\widehat{b}_{m,S}\|_S^2$ by

$$\|\widehat{b}_m\|_S^2 \le (c_S^\star)^2 \frac{n}{(\log(n))^2}\|Y\|^2.$$

Then we just need to prove that $\mathbb{P}(\Omega_{m,S}^c) \le cn^{-p}$.

The lines of the proof of Theorem 1 can also be followed with (16) which writes now

$$\|t\|_{n,S}^2 = Z_n(t) + \frac{1}{n}\left\{\mathbb{E}[t(X_1)]\right\}^2 + \frac{2}{n}\mathbb{E}[t(X_1)]V_n(t) - (1 - \frac{1}{n})U_n(t) \tag{45}$$

where $Z_n(t), V_n(t)$ and $U_n(t)$ are unchanged. As a consequence, we still rely on $\Omega_m^{(1)}$ and $\Omega_m^{(2)}$ and Lemmas 3 and 4, leading to $\Omega_m^{(1)} \cap \Omega_m^{(2)} \subset \Omega_{m,S}$. □

Proof of Lemma 2. The proof of the Lemma follows the line of the proof of Proposition 3, where we replace α_n by 1 and c_n by $1/n$, as $\Psi_m^{(S)} = \Psi_m + \frac{1}{n}\Xi_m$. □

9. Appendix

9.1. **Properties of some Matrices.** We state properties of matrices $\Sigma_n(a,b) = (a_{ij})_{1\le i,j\le n}$ such that $a_{ii} = a$ for $i = 1,\dots,n$ and $a_{ij} = b$ for $1 \le i \neq j \le n$. The matrix $\Sigma_n(a,b)$ is symmetric and

$$\Sigma_n(a,b) - (a-b)I_n = \Sigma_n(b,b)$$

has rank 1, while $\Sigma_n(a,b)\mathbf{1} = (a + (n-1)b)\mathbf{1}$ Thus, $(a-b)$ is eigenvalue of order $n-1$ and $a + (n-1)b$ is the other eigenvalue. Therefore, $\Sigma_n(a,b)$ is positive definite if and only if

$$a - b > 0, \quad a + (n-1)b > 0. \tag{46}$$

We obtain $\Sigma_n(a,b)^{-1}$ by solving the system $\Sigma_n(a,b)\mathbf{x} = \mathbf{y}$ and find $\Sigma_n(a,b)^{-1} = \Sigma_n(\alpha,\beta)$ with

$$\alpha = \frac{a + (n-2)b}{(a-b)[a + (n-1)b]}, \quad \beta = -\frac{b}{(a-b)[a+(n-1)b]} \tag{47}$$

The eigenvalues of $\Sigma_n(a,b)^{-1}$ are $(a-b)^{-1}, (a+(n-1)b)^{-1}$.

The product of two matrices $\Sigma_n(a,b)$ and $\Sigma_n(a',b')$ is equal to $\Sigma_n(A,B)$ with $A = aa' + (n-1)bb', B = ab' + a'b + (n-2)bb'$.

All matrices $\Sigma_n(a,b)$ can be diagonalized using the same eigenbasis: $\epsilon_1 = (1/\sqrt{n})\mathbf{1}$, and $(\epsilon_i)_{2\le i\le n}$ an orthonormal basis of $(\mathrm{Vect}(\epsilon_1))^\perp$.

9.2. **Useful result.** A proof of the following theorem can be found in Stewart and Sun (1990).

Theorem 4. *Let* $\mathbf{A}$, $\mathbf{B}$ *be* $(m \times m)$ *matrices. If* $\mathbf{A}$ *is invertible and* $\|\mathbf{A}^{-1}\mathbf{B}\|_{\mathrm{op}} < 1$, *then* $\tilde{\mathbf{A}} := \mathbf{A} + \mathbf{B}$ *is invertible and it holds*

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

References

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

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

[3] Barron, A., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113**, 301-413. *The Annals of the Institute of Statistical Mathematics*, **71**, 29-62.

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

[5] Carmona, R. Delarue, F. and Lacker, D. (2016). Mean field games with common noise. *Ann. Probab.*, **44** (6), 3740-3803.

[6] Caron, E., Dedecker, J., Michel, B., (2021). Gaussian linear model selection in a dependent context. *Electron. J. Stat.* **15**, no. 2, 4823-4867.

[7] 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.

[8] Cohen, A., Davenport, M.A. and Leviatan, D. (2019). Correction to: On the Stability and Accuracy of Least Squares Approximations. Foundations of Computational Mathematics **19**, 239.

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

[10] Comte, F. and Genon-Catalot, V. (2019). Regression function estimation on non compact support as a partly inverse problem. F. Comte, V. Genon-Catalot. *Annals of the Institute of Statistical Mathematics* **72**, 1023-1054.

[11] Comte, F. and Lacour, C. (2023). Noncompact estimation of the conditional density from direct or noisy data. *Annals of the Institute H. Poincaré, Probability and Statistics*, **59**, 1463-1507.

[12] Delarue, F., Tanré, E. and Maillet R. (2024). Ergodicity of some stochastic Fokker-Planck equations with additive common noise. *arXiv preprint.* arXiv:2405.09950.

[13] Genon-Catalot, V. and Larédo, C. (2025). LAMN property for drift inference in systems of stochastic differential equations with common noise. *Preprint Hal* 05027190.

[14] Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. Giné, Evariste (ed.) et al., *Stochastic inequalities and applications.* Selected papers presented at the Euroconference on "Stochastic inequalities and their applications", Barcelona, June 18–22, 2002. Basel: Birkhäuser. Prog. Probab. 56, 55-69.

[15] Lacker, D., Shkolnikov, M. and Zhang J. (2022). Superposition and mimicking theorems for conditional McKean-Vlasov equations. *Journal of the European Mathematical Society*, **25** (8), 3229-3288.

[16] Maillet, R. (2025). A note on the Long-Time behaviour of Stochastic McKean-Vlasov Equations with common noise. *arXiv preprint* arXiv:2306.16130.

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

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

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

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

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

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