Model selection and estimation of a component in additive regression

Xavier Gendre

Model selection and estimation of a component in additive regression

Xavier Gendre

Institut de Mathématiques de Toulouse
Université de Toulouse et CNRS (UMR 5219)

Xavier.Gendre@math.univ-toulouse.fr

Abstract

Let $Y \in \mathbb{R}^n$ be a random vector with mean s and covariance matrix $\sigma^2 P_n {}^tP_n$ where P_n is some known $n \times n$-matrix. We construct a statistical procedure to estimate s as well as under moment condition on Y or Gaussian hypothesis. Both cases are developed for known or unknown σ^2. Our approach is free from any prior assumption on s and is based on non-asymptotic model selection methods. Given some linear spaces collection $\{S_m,\ m \in \mathcal{M}\}$, we consider, for any $m \in \mathcal{M}$, the least-squares estimator $\hat{s}_m$ of s in S_m. Considering a penalty function that is not linear in the dimensions of the S_m's, we select some $\hat{m} \in \mathcal{M}$ in order to get an estimator $\hat{s}_{\hat{m}}$ with a quadratic risk as close as possible to the minimal one among the risks of the $\hat{s}_m$'s. Non-asymptotic oracle-type inequalities and minimax convergence rates are proved for $\hat{s}_{\hat{m}}$. A special attention is given to the estimation of a non-parametric component in additive models. Finally, we carry out a simulation study in order to illustrate the performances of our estimators in practice.

1 Introduction

1.1 Additive models

The general form of a *regression model* can be expressed as

$$Z = f(X) + \sigma\varepsilon \tag{1}$$

where $X = (X^{(1)}, \dots, X^{(k)})'$ is the k-dimensional vector of *explanatory variables* that belongs to some product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k \subset \mathbb{R}^k$, the unknown function $f : \mathcal{X} \to \mathbb{R}$ is called *regression function*, the positive real number σ is a standard deviation factor and the real random noise ε is such that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] < \infty$ almost surely.

In such a model, we are interested in the behavior of Z in accordance with the fluctuations of X. In other words, we want to explain the random variable Z through the function $f(x) = \mathbb{E}[Z|X = x]$. For this purpose, many approaches have been proposed and, among them, a widely used is the *linear regression*

$$Z = \mu + \sum_{i=1}^{k} \beta_i X^{(i)} + \sigma\varepsilon \tag{2}$$

where μ and the β_i's are unknown constants. This model benefits from easy interpretation in practice and, from a statistical point of view, allows componentwise analysis. However, a

drawback of linear regression is its lack of flexibility for modeling more complex dependencies between Z and the $X^{(i)}$'s. In order to bypass this problem while keeping the advantages of models like (2), we can generalize them by considering *additive regression models* of the form

$$Z = \mu + \sum_{i=1}^{k} f_i(X^{(i)}) + \sigma\varepsilon \tag{3}$$

where the unknown functions $f_i : \mathcal{X}_i \to \mathbb{R}$ will be referred to as the *components* of the regression function f. The object of this paper is to construct a data-driven procedure for estimating one of these components on a fixed design (*i.e.* conditionally to some realizations of the random variable X). Our approach is based on nonasymptotic model selection and is free from any prior assumption on f and its components. In particular, we do not make any regularity hypothesis on the function to estimate except to deduce uniform convergence rates for our estimators.

Models (3) are not new and were first considered in the context of input-output analysis by Leontief [23] and in analysis of variance by Scheffé [35]. This kind of model structure is widely used in theoretical economics and in econometric data analysis and leads to many well known economic results. For more details about interpretability of additive models in economics, the interested reader could find many references at the end of Chapter 8 of [18].

As we mention above, regression models are useful for interpreting the effects of X on changes of Z. To this end, the statisticians have to estimate the regression function f. Assuming that we observe a sample $\{(X_1, Z_1), \ldots, (X_n, Z_n)\}$ obtained from model (1), it is well known (see [37]) that the optimal $\mathbb{L}^2$ convergence rate for estimating f is of order $n^{-\alpha/(2\alpha+k)}$ where $\alpha > 0$ is an index of smoothness of f. Note that, for large value of k, this rate becomes slow and the performances of any estimation procedure suffer from what is called the *curse of the dimension* in literature. In this connection, Stone [37] has proved the notable fact that, for additive models (3), the optimal $\mathbb{L}^2$ convergence rate for estimating each component f_i of f is the one-dimensional rate $n^{-\alpha/(2\alpha+1)}$. In other terms, estimation of the component f_i in (3) can be done with the same optimal rate than the one achievable with the model $Z' = f_i(X^{(i)}) + \sigma\varepsilon$.

Components estimation in additive models has received a large interest since the eighties and this theory benefited a lot from the the works of Buja *et al.* [15], Hastie and Tibshirani [19]. Very popular methods for estimating components in (3) are based on *backfitting* procedures (see [12] for more details). These techniques are iterative and may depend on the starting values. The performances of these methods deeply depends on the choice of some convergence criterion and the nature of the obtained results is usually asymptotic (see, for example, the works of Opsomer and Ruppert [30] and Mammen, Linton and Nielsen [26]). More recent non-iterative methods have been proposed for estimating marginal effects of the $X^{(i)}$ on the variable Z (*i.e.* how Z fluctuates on average if one explanatory variable is varying while others stay fixed). These procedures, known as *marginal integration estimation*, were introduced by Tjøstheim and Auestad [38] and Linton and Nielsen [24]. In order to estimate the marginal effect of $X^{(i)}$, these methods take place in two times. First, they estimate the regression function f by a particular estimator f^*, called *pre-smoother*, and then they average f^* according to all the variables except $X^{(i)}$. The way for constructing f^* is fundamental and, in practice, one uses a special kernel estimator (see [34] and [36] for a discussion on this subject). To this end, one needs to estimate two unknown bandwidths that are necessary for getting f^*. Dealing with a finite sample, the impact of how we estimate these bandwidths is

not clear and, as for backfitting, the theoretical results obtained by these methods are mainly asymptotic.

In contrast with these methods, we are interested here in nonasymptotic procedures to estimate components in additive models. The following subsection is devoted to introduce some notations and the framework that we handle but also a short review of existing results in nonasymptotic estimation in additive models.

1.2 Statistical framework

We are interested in estimating one of the components in the model (3) with, for any i, $\mathcal{X}_i = [0,1]$. To focus on it, we denote by $s : [0,1] \to \mathbb{R}$ the component that we plan to estimate and by $t^1, \ldots, t^K : [0,1] \to \mathbb{R}$ the $K \geqslant 1$ other ones. Thus, considering the design points $(x_1, y_1^1, \ldots, y_1^K)', \ldots, (x_1, y_1^1, \ldots, y_1^K)' \in [0,1]^{K+1}$, we observe

$$Z_i = s(x_i) + \mu + \sum_{j=1}^{K} t^j(y_i^j) + \sigma\varepsilon_i, \;\; i = 1, \ldots, n \;, \tag{4}$$

where the components $s, t^1, \ldots, t^K$ are unknown functions, μ in an unknown real number, σ is a positive factor and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unobservable centered random vector with i.i.d. components of unit variance.

Let ν be a probability measure on $[0,1]$, we introduce the space of centered and square-integrable functions

$$\mathbb{L}_0^2([0,1], \nu) = \left\{ f \in \mathbb{L}^2([0,1], \nu) \;\; : \;\; \int_0^1 f(t)\nu(dt) = 0 \right\} \;.$$

Let $\nu_1, \ldots, \nu_K$ be K probability measures on $[0,1]$, to avoid identification problems in the sequel, we assume

$$s \in \mathbb{L}_0^2\left([0,1], \nu\right) \;\text{ and }\; t^j \in \mathbb{L}_0^2\left([0,1], \nu_j\right), \;\; j = 1, \ldots, K \;. \tag{5}$$

This hypothesis is not restrictive since we are interested in how $Z = (Z_1, \ldots, Z_n)'$ fluctuates with respect to the x_i's. A shift on the components does not affect these fluctuations and the estimation proceeds up to the additive constant μ.

The results described in this paper are obtained under two different assumptions on the noise terms ε_i, namely

$(\mathbf{H_{Gau}})$ the random vector ε is a standard Gaussian vector in $\mathbb{R}^n$,

and

$(\mathbf{H_{Mom}})$ the variables ε_i satisfy the moment condition

$$\exists p > 2 \text{ such that } \forall i, \;\; \tau_p = \mathbb{E}\left[|\varepsilon_i|^p\right] < \infty \;. \tag{6}$$

Obviously, **$(\mathbf{H_{Mom}})$** is weaker than **$(\mathbf{H_{Gau}})$**. We consider these two cases in order to illustrate how better are the results in the Gaussian case with regard to the moment condition case. From the point of view of model selection, we show in the corollaries of Section 2 that we are

allowed to work with more general model collections under $(\mathbf{H_{Gau}})$ than under $(\mathbf{H_{Mom}})$ in order to get similar results. Thus, the main contribution of the Gaussian assumption is to give more flexibility to the procedure described in the sequel.

So, our aim is to estimate the component s on the basis of the observations (4). For the sake of simplicity of this introduction, we assume that the quantity $\sigma^2 > 0$ is known (see Section 3 for unknown variance) and we introduce the vectors $s = (s_1, \ldots, s_n)'$ and $t = (t_1, \ldots, t_n)'$ defined by, for any $i \in \{1, \ldots, n\}$,

$$s_i = s(x_i) \quad \text{and} \quad t_i = \mu + \sum_{j=1}^{K} t^j(y_i^j) \ . \tag{7}$$

Moreover, we assume that we know two linear subspaces $E, F \subset \mathbb{R}^n$ such that $s \in E$, $t \in F$ and $E \oplus F = \mathbb{R}^n$. Of course, such spaces are not available to the statisticians in practice and, when we handle additive models in Section 4, we will not suppose that they are known. Let P_n be the projection onto E along F, we derive from (4) the following regression framework

$$Y = P_n Z = s + \sigma P_n \varepsilon \tag{8}$$

where $Y = (Y_1, \ldots, Y_n)'$ belongs to $E = \mathrm{Im}(P_n) \subset \mathbb{R}^n$.

The framework (8) is similar to the classical *signal-plus-noise* regression framework but the data are not independent and their variances are not equal. Because of this uncommonness of the variances of the observations, we qualify (8) as an *heteroscedastic* framework. The object of this paper is to estimate the component s and we handle (8) to this end. The particular case of P_n equal to the unit matrix has been widely treated in the literature (see, for example, [10] for $(\mathbf{H_{Gau}})$ and [4] for $(\mathbf{H_{Mom}})$). The case of an unknown but diagonal matrix P_n has been studied in several papers for the Gaussian case (see, for example, [16] and [17]). By using cross-validation and resampling penalties, Arlot and Massart [3] and Arlot [2] have also considered the framework (8) with unknown diagonal matrix P_n. Laurent, Loubes and Marteau [21] deal with a known diagonal matrix P_n for studying testing procedure in an inverse problem framework. The general case of a known non-diagonal matrix P_n naturally appears in applied fields as, for example, genomic studies (see Chapters 4 and 5 of [33]).

The results that we introduce in the sequel consider the framework (8) from a general outlook and we do not make any prior hypothesis on P_n. In particular, we do not suppose that P_n is invertible. We only assume that it is a projector when we handle the problem of component estimation in an additive framework in Section 4. Without loss of generality, we always admit that $s \in \mathrm{Im}(P_n)$. Indeed, if s does not belong to $\mathrm{Im}(P_n)$, it suffices to consider the orthogonal projection π_{P_n} onto $\mathrm{Im}(P_n)^\perp$ and to notice that $\pi_{P_n} Y = \pi_{P_n} s$ is not random. Thus, replacing Y by $Y - \pi_{P_n} Y$ leads to (8) with a mean lying in $\mathrm{Im}(P_n)$. For general matrix P_n, other approaches could be used. However, for the sake of legibility, we consider $s \in \mathrm{Im}(P_n)$ because, for the estimation of a component in an additive framework, by construction, we always have $Y = P_n Z \in \mathrm{Im}(P_n)$ as it will be specified in Section 4.

We now describe our estimation procedure in details. For any $z \in \mathbb{R}^n$, we define the *least-squares contrast* by

$$\gamma_n(z) = \|Y - z\|_n^2 = \frac{1}{n} \sum_{i=0}^{n} (Y_i - z_i)^2 \ .$$

Let us consider a collection of linear subspaces of $\mathrm{Im}(P_n)$ denoted by $\mathcal{F} = \{S_m, \ m \in \mathcal{M}\}$ where $\mathcal{M}$ is a finite or countable index set. Hereafter, the S_m's will be called the *models*.

Denoting by π_m the orthogonal projection onto S_m, the minimum of γ_n over S_m is achieved at a single point $\hat{s}_m = \pi_m Y$ called the *least-squares estimator* of s in S_m. Note that the expectation of $\hat{s}_m$ is equal to the orthogonal projection $s_m = \pi_m s$ of s onto S_m. We have the following identity for the quadratic risks of the $\hat{s}_m$'s,

Proposition 1.1. *Let $m \in \mathcal{M}$, the least-squares estimator $\hat{s}_m = \pi_m Y$ of s in S_m satisfies*

$$\mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] = \|s - s_m\|_n^2 + \frac{Tr({}^tP_n \pi_m P_n)}{n}\sigma^2 \tag{9}$$

where $Tr(\cdot)$ is the trace operator.

Proof. By orthogonality, we have

$$\|s - \hat{s}_m\|_n^2 = \|s - s_m\|_n^2 + \sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \ . \tag{10}$$

Because the components of ε are independent and centered with unit variance, we easily compute

$$\mathbb{E}\left[\|\pi_m P_n \varepsilon\|_n^2\right] = \frac{\mathrm{Tr}({}^tP_n \pi_m P_n)}{n} \ .$$

We conclude by taking the expectation on both side of (10). □

A "good" estimator is such that its quadratic risk is small. The decomposition given by (9) shows that this risk is a sum of two non-negative terms that can be interpreted as follows. The first one, called *bias term*, corresponds to the capacity of the model S_m to approximate the true value of s. The second, called *variance term*, is proportional to $\mathrm{Tr}({}^tP_n \pi_m P_n)$ and measures, in a certain sense, the complexity of S_m. If $S_m = \mathbb{R}u$, for some $u \in \mathbb{R}^n$, then the variance term is small but the bias term is as large as s is far from the too simple model S_m. Conversely, if S_m is a "huge" model, whole $\mathbb{R}^n$ for instance, the bias is null but the price is a great variance term. Thus, (9) illustrates why choosing a "good" model amounts to finding a trade-off between bias and variance terms.

Clearly, the choice of a model that minimizes the risk (9) depends on the unknown vector s and makes good models unavailable to the statisticians. So, we need a data-driven procedure to select an index $\hat{m} \in \mathcal{M}$ such that $\mathbb{E}[\|s - \hat{s}_{\hat{m}}\|_n^2]$ is close to the smaller $\mathbb{L}^2$-risk among the collection of estimators $\{\hat{s}_m,\ m \in \mathcal{M}\}$, namely

$$\mathcal{R}(s, \mathcal{F}) = \inf_{m \in \mathcal{M}} \mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] \ .$$

To choose such a $\hat{m}$, a classical way in model selection consists in minimizing an empirical penalized criterion stochastically close to the risk. Given a *penalty* function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$, we define $\hat{m}$ as any minimizer over $\mathcal{M}$ of the penalized least-squares criterion

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\} \ . \tag{11}$$

This way, we select a model $S_{\hat{m}}$ and we have at our disposal the *penalized least-squares estimator* $\tilde{s} = \hat{s}_{\hat{m}}$. Note that, by definition, the estimator $\tilde{s}$ satisfies

$$\forall m \in \mathcal{M},\ \gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) \leqslant \gamma_n(\hat{s}_m) + \mathrm{pen}(m) \ . \tag{12}$$

To study the performances of $\tilde{s}$, we have in mind to upperbound its quadratic risk. To this end, we establish inequalities of the form

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C \inf_{m\in\mathcal{M}} \left\{\|s-s_m\|_n^2 + \mathrm{pen}(m)\right\} + \frac{R}{n} \tag{13}$$

where C and R are numerical terms that do not depend on n. Note that if the penalty is proportional to $\mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n$, then the quantity involved in the infimum is of order of the $\mathbb{L}^2$-risk of $\hat{s}_m$. Consequently, under suitable assumptions, such inequalities allow us to deduce upperbounds of order of the minimal risk among the collection of estimators $\{\hat{s}_m,\ m\in\mathcal{M}\}$. This result is known as an *oracle inequality*

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C\mathcal{R}(s,\mathcal{F}) = C \inf_{m\in\mathcal{M}} \mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right] \ . \tag{14}$$

This kind of procedure is not new and the first results in estimation by penalized criterion are due to Akaike [1] and Mallows [25] in the early seventies. Since these works, model selection has known an important development and it would be beyond the scope of this paper to make an exhaustive historical review of the domain. We refer to the first chapters of [28] for a more general introduction.

Nonasymptotic model selection approach for estimating components in an additive model was studied in few papers only. Considering penalties that are linear in the dimension of the models, Baraud, Comte and Viennet [6] have obtained general results for geometrically β-mixing regression models. Applying it to the particular case of additive models, they estimate the whole regression function. They obtain nonasymptotic upperbounds similar to (13) on condition ε admits a moment of order larger than 6. For additive regression on a random design and alike penalties, Baraud [5] proved oracle inequalities for estimators of the whole regression function constructed with polynomial collections of models and a noise that admits a moment of order 4. Recently, Brunel and Comte [13] have obtained results with the same flavor for the estimation of the regression function in a censored additive model and a noise admitting a moment of order larger than 8. Pursuant to this work, Brunel and Comte [14] have also proposed a nonasymptotic iterative method to achieve the same goal. Combining ideas from sparse linear modeling and additive regression, Ravikumar *et al.* [32] have recently developed a data-driven procedure, called SpAM, for estimating a sparse high-dimensional regression function. Some of their empirical results have been proved by Meier, van de Geer and Bühlmann [29] in the case of a sub-Gaussian noise and some sparsity-smoothness penalty.

The methods that we use are similar to the ones of Baraud, Comte and Viennet and are inspired from [4]. The main contribution of this paper is the generalization of the results of [4] and [6] to the framework (8) with a known matrix P_n under Gaussian hypothesis or only moment condition on the noise terms. Taking into account the correlations between the observations in the procedure leads us to deal with penalties that are not linear in the dimension of the models. Such a consideration naturally arises in heteroscedastic framework. Indeed, as mentioned in [2], at least from an asymptotic point of view, considering penalties linear in the dimension of the models in an heteroscedastic framework does not lead to oracle inequalities for $\tilde{s}$. For our penalized procedure and under mild assumptions on $\mathcal{F}$, we prove oracle inequalities under Gaussian hypothesis on the noise or only under some moment condition.

Moreover, we introduce a nonasymptotic procedure to estimate one component in an additive framework. Indeed, the works cited above are all connected to the estimation of the whole regression function by estimating simultaneously all of its components. Since these components are each treated in the same way, their procedures can not focus on the properties of

one of them. In the procedure that we propose, we can be sharper, from the point of view of the bias term, by using more models to estimate a particular component. This allows us to deduce uniform convergence rates over Hölderian balls and adaptivity of our estimators. Up to the best of our knowledge, our results in nonasymptotic estimation of a nonparametric component in an additive regression model are new.

The paper is organized as follows. In Section 2, we study the properties of the estimation procedure under the hypotheses $(\mathbf{H_{Gau}})$ and $(\mathbf{H_{Mom}})$ with a known variance factor σ^2. As a consequence, we deduce oracle inequalities and we discuss about the size of the collection $\mathcal{F}$. The case of unknown σ^2 is presented in Section 3 and the results of the previous section are extended to this situation. In Section 4, we apply these results to the particular case of the additive models and, in the next section, we give uniform convergence rates for our estimators over Hölderian balls. Finally, in Section 6, we illustrate the performances of our estimators in practice by a simulation study. The last sections are devoted to the proofs and to some technical lemmas.

Notations: in the sequel, for any $x = (x_1, \ldots, x_n)'$, $y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$, we define

$$\|x\|_n^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 \quad \text{and} \quad \langle x, y\rangle_n = \frac{1}{n}\sum_{i=1}^n x_i y_i \ .$$

We denote by ρ the *spectral norm* on the set $\mathbb{M}_n$ of the $n \times n$ real matrices as the norm induced by $\|\cdot\|_n$,

$$\forall A \in \mathbb{M}_n, \ \rho(A) = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_n}{\|x\|_n} \ .$$

For more details about the properties of ρ, see Chapter 5 of [20].

2 Main results

Throughout this section, we deal with the statistical framework given by (8) with $s \in \mathrm{Im}(P_n)$ and we assume that the variance factor σ^2 is known. Moreover, in the sequel of this paper, for any $d \in \mathbb{N}$, we define N_d as the number of models of dimension d in $\mathcal{F}$,

$$N_d = \mathrm{Card}\left\{m \in \mathcal{M} \ : \ \dim(S_m) = d\right\} \ .$$

We first introduce general model selection theorems under hypotheses $(\mathbf{H_{Gau}})$ and $(\mathbf{H_{Mom}})$.

Theorem 2.1. *Assume that* $(\mathbf{H_{Gau}})$ *holds and consider a collection of nonnegative numbers* $\{L_m, m \in \mathcal{M}\}$. *Let* $\theta > 0$, *if the penalty function is such that*

$$pen(m) \geqslant (1 + \theta + L_m)\frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2 \ \text{ for all } m \in \mathcal{M} \ , \tag{15}$$

then the penalized least-squares estimator $\tilde{s}$ *given by (11) satisfies*

$$\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right] \leqslant C \inf_{m \in \mathcal{M}} \left\{\|s - s_m\|_n^2 + pen(m) - \frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2\right\} + \frac{\rho^2(P_n)\sigma^2}{n}R_n(\theta) \tag{16}$$

where we have set

$$R_n(\theta) = C' \sum_{m \in \mathcal{M}} \exp\left(-\frac{C'' L_m \, Tr({}^tP_n\pi_m P_n)}{\rho^2(P_n)}\right)$$

and $C > 1$ *and* $C', C'' > 0$ *are constants that only depend on* θ.

If the errors are not supposed to be Gaussian but only to satisfy the moment condition $(\mathbf{H_{Mom}})$, the following upperbound on the q-th moment of $\|s-\tilde{s}\|_n^2$ holds.

Theorem 2.2. *Assume that* $(\mathbf{H_{Mom}})$ *holds and take* $q>0$ *such that* $2(q+1)<p$. *Consider* $\theta>0$ *and some collection* $\{L_m,\ m\in\mathcal{M}\}$ *of positive weights. If the penalty function is such that*

$$pen(m) \geqslant (1+\theta+L_m)\frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2 \text{ for all } m\in\mathcal{M}\ , \tag{17}$$

then the penalized least-squares estimator $\tilde{s}$ *given by (11) satisfies*

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]^{1/q} \leqslant C \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2+pen(m)\right\}+\frac{\rho^2(P_n)\sigma^2}{n}R_n(p,q,\theta)^{1/q} \tag{18}$$

where we have set $R_n(p,q,\theta)$ *equal to*

$$C'\tau_p\left[N_0+\sum_{m\in\mathcal{M}:S_m\neq\{0\}}\left(1+\frac{Tr({}^tP_n\pi_m P_n)}{\rho^2(\pi_m P_n)}\right)\left(\frac{L_m\,Tr({}^tP_n\pi_m P_n)}{\rho^2(P_n)}\right)^{q-p/2}\right]$$

and $C=C(q,\theta)$, $C'=C'(p,q,\theta)$ *are positive constants.*

The proofs of these theorems give explicit values for the constants C that appear in the upperbounds. In both cases, these constants go to infinity as θ tends to 0 or increases toward infinity. In practice, it does neither seem reasonable to choose θ close to 0 nor very large. Thus this explosive behavior is not restrictive but we still have to choose a "good" θ. The values for θ suggested by the proofs are around the unity but we make no claim of optimality. Indeed, this is a hard problem to determine an optimal choice for θ from theoretical computations since it could depend on all the parameters and on the choice of the collection of models. In order to calibrate it in practice, several solutions are conceivable. We can use a simulation study, deal with cross-validation or try to adapt the slope heuristics described in [11] to our procedure.

For penalties of order of $\mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n$, Inequalities (16) and (18) are not far from being oracle. Let us denote by R_n the remainder term $R_n(\theta)$ or $R_n(p,q,\theta)$ according to whether $(\mathbf{H_{Gau}})$ or $(\mathbf{H_{Mom}})$ holds. To deduce oracle inequalities from that, we need some additional hypotheses as the following ones:

$(\mathbf{A}_1)$ there exists some universal constant $\zeta>0$ such that

$$\mathrm{pen}(m) \leqslant \zeta\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2, \text{ for all } m\in\mathcal{M}\ ,$$

$(\mathbf{A}_2)$ there exists some constant $R>0$ such that

$$\sup_{n\geqslant 1} R_n \leqslant R\ ,$$

$(\mathbf{A}_3)$ there exists some constant $\rho>1$ such that

$$\sup_{n\geqslant 1}\rho^2(P_n)\leqslant\rho^2\ .$$

Thus, under the hypotheses of Theorem 2.1 and these three assumptions, we deduce from (16) that

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + \frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2\right\} + \frac{R\rho^2\sigma^2}{n}$$

where C is a constant that does not depend on s, σ^2 and n. By Proposition 1.1, this inequality corresponds to (14) up to some additive term. To derive similar inequality from (18), we need on top of that to assume that $p > 4$ in order to be able to take $q = 1$.

Assumption $\mathbf{(A_3)}$ is subtle and strongly depends on the nature of P_n. The case of oblique projector that we use to estimate a component in an additive framework will be discussed in Section 4. Let us replace it, for the moment, by the following one

$\mathbf{(A_3')}$ there exists $c \in (0,1)$ that does not depend on n such that

$$c\rho^2(P_n)\dim(S_m) \leqslant \mathrm{Tr}({}^tP_n\pi_m P_n)\ .$$

By the properties of the norm ρ, note that $\mathrm{Tr}({}^tP_n\pi_m P_n)$ always admits an upperbound with the same flavor

$$\begin{aligned}\mathrm{Tr}({}^tP_n\pi_m P_n) &= \mathrm{Tr}(\pi_m P_n\, {}^t(\pi_m P_n))\\ &\leqslant \rho(\pi_m P_n\, {}^t(\pi_m P_n))\mathrm{rk}(\pi_m P_n\, {}^t(\pi_m P_n))\\ &\leqslant \rho^2(\pi_m P_n)\mathrm{rk}(\pi_m)\\ &\leqslant \rho^2(P_n)\dim(S_m)\ .\end{aligned}$$

In all our results, the quantity $\mathrm{Tr}({}^tP_n\pi_m P_n)$ stands for a dimensional term relative to S_m. Hypothesis $\mathbf{(A_3')}$ formalizes that by assuming that its order is the dimension of the model S_m up to the norm of the covariance matrix ${}^tP_nP_n$.

Let us now discuss about the assumptions $\mathbf{(A_1)}$ and $\mathbf{(A_2)}$. They are connected and they raise the impact of the complexity of the collection $\mathcal{F}$ on the estimation procedure. Typically, condition $\mathbf{(A_2)}$ will be fulfilled under $\mathbf{(A_1)}$ when $\mathcal{F}$ is not too "large", that is, when the collection does not contain too many models with the same dimension. We illustrate this phenomenon by the two following corollaries.

Corollary 2.1. *Assume that* $\mathbf{(H_{Gau})}$ *and* $\mathbf{(A_3')}$ *hold and consider some finite* $A \geqslant 0$ *such that*

$$\sup_{d\in\mathbb{N}:N_d>0}\frac{\log N_d}{d} \leqslant A\ . \tag{19}$$

Let L, θ *and* ω *be some positive numbers that satisfy*

$$L \geqslant \frac{2(1+\theta)^3}{c\theta^2}(A+\omega)\ .$$

Then, the estimator $\tilde{s}$ *obtained from (11) with penalty function given by*

$$pen(m) = (1+\theta+L)\frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2$$

is such that

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (L\vee 1)\frac{Tr({}^tP_n\pi_m P_n)\vee(c\rho^2(P_n))}{n}\sigma^2\right\}$$

where $C > 1$ *only depends on* θ, ω *and* c.

For errors that only satisfy moment condition, we have the following similar result.

Corollary 2.2. *Assume that* $(\mathbf{H_{Mom}})$ *and* $(\mathbf{A}'_3)$ *hold with* $p > 6$ *and let* $A > 0$ *and* $\omega > 0$ *such that*

$$N_0 \leqslant 1 \quad \text{and} \quad \sup_{d>0:N_d>0} \frac{N_d}{(1+d)^{p/2-3-\omega}} \leqslant A \ . \tag{20}$$

Consider some positive numbers L, θ *and* ω' *that satisfy*

$$L \geqslant \omega' A^{2/(p-2)} \ ,$$

then, the estimator $\tilde{s}$ *obtained from (11) with penalty function given by*

$$pen(m) = (1+\theta+L)\frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2$$

is such that

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C\tau_p \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (L\vee 1)\frac{Tr({}^tP_n\pi_m P_n)\vee(c\rho^2(P_n))}{n}\sigma^2\right\}$$

where $C > 1$ *only depends on* θ, p, ω, ω' *and* c.

Note that the assumption $(\mathbf{A}'_3)$ guarantees that $\mathrm{Tr}({}^tP_n\pi_m P_n)$ is not smaller than $c\rho^2(P_n)\dim(S_m)$ and, at least for the models with positive dimension, this implies $\mathrm{Tr}({}^tP_n\pi_m P_n) \geqslant c\rho^2(P_n)$. Consequently, up to the factor L, the upperbounds of $\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right]$ given by Corollaries 2.1 and 2.2 are of order of the minimal risk $\mathcal{R}(s,\mathcal{F})$. To deduce oracle inequalities for $\tilde{s}$ from that, $(\mathbf{A}_1)$ needs to be fulfilled. In other terms, we need to be able to consider some L independently from the size n of the data. It will be the case if the same is true for the bounds A.

Let us assume that the collection $\mathcal{F}$ is small in the sense that, for any $d \in \mathbb{N}$, the number of models N_d is bounded by some constant term that neither depends on n nor d. Typically, collections of nested models satisfy that. In this case, we are free to take L equal to some universal constant. So, $(\mathbf{A}_1)$ is true for $\zeta = 1+\theta+L$ and oracle inequalities can be deduced for $\tilde{s}$. Conversely, a large collection $\mathcal{F}$ is such that there are many models with the same dimension. We consider that this situation happens, for example, when the order of A is $\log n$. In such a case, we need to choose L of order $\log n$ too and the upperbounds on the risk of $\tilde{s}$ become oracle type inequalities up to some logarithmic factor. However, we know that in some situations, this factor can not be avoided as in the complete variable selection problem with Gaussian errors (see Chapter 4 of [27]).

As a consequence, the same model selection procedure allows us to deduce oracle type inequalities under $(\mathbf{H_{Gau}})$ and $(\mathbf{H_{Mom}})$. Nevertheless, the assumption on N_d in Corollary 2.2 is more restrictive than the one in Corollary 2.1. Indeed, to obtain an oracle inequality in the Gaussian case, the quantity N_d is limited by e^{Ad} while the bound is only polynomial in d under moment condition. Thus, the Gaussian assumption $(\mathbf{H_{Gau}})$ allows to obtain oracle inequalities for more general collections of models.

3 Estimation when variance is unknown

In contrast with Section 2, the variance factor σ^2 is here assumed to be unknown in (8). Since the penalties given by Theorems 2.1 and 2.2 depend on σ^2, the procedure introduced in the

previous section does not remain available to the statisticians. Thus, we need to estimate σ^2 in order to replace it in the penalty functions. The results of this section give upperbounds for the $\mathbb{L}^2$-risk of the estimators $\tilde{s}$ constructed in such a way.

To estimate the variance factor, we use a residual least-squares estimator $\hat{\sigma}^2$ that we define as follows. Let V be some linear subspace of $\mathrm{Im}(P_n)$ such that

$$\mathrm{Tr}({}^tP_n\pi P_n) \leqslant \mathrm{Tr}({}^tP_nP_n)/2 \tag{21}$$

where π is the orthogonal projection onto V. We define

$$\hat{\sigma}^2 = \frac{n\|Y-\pi Y\|_n^2}{\mathrm{Tr}\left({}^tP_n(I_n-\pi)P_n\right)} \ . \tag{22}$$

First, we assume that the errors are Gaussian. The following result holds.

Theorem 3.1. *Assume that* $\mathbf{(H_{Gau})}$ *holds. For any* $\theta>0$*, we define the penalty function*

$$\forall m\in\mathcal{M},\ \ pen(m) = (1+\theta)\frac{Tr({}^tP_n\pi_mP_n)}{n}\hat{\sigma}^2 \ . \tag{23}$$

Then, for some positive constants C*,* C' *and* C'' *that only depend on* θ*, the penalized least-squares estimator* $\tilde{s}$ *satisfies*

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C\left(\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right]+\|s-\pi s\|_n^2\right)+\frac{\rho^2(P_n)\sigma^2}{n}\bar{R}_n(\theta) \tag{24}$$

where we have set

$$\bar{R}_n(\theta) = C'\left[\left(2+\frac{\|s\|_n^2}{\rho^2(P_n)\sigma^2}\right)\exp\left(-\frac{\theta^2\,Tr({}^tP_nP_n)}{32\rho^2(P_n)}\right)+\sum_{m\in\mathcal{M}}\exp\left(-C''\frac{Tr({}^tP_n\pi_mP_n)}{\rho^2(P_n)}\right)\right] \ .$$

If the errors are only assumed to satisfy a moment condition, we have the following theorem.

Theorem 3.2. *Assume that* $\mathbf{(H_{Mom})}$ *holds. Let* $\theta>0$*, we consider the penalty function defined by*

$$\forall m\in\mathcal{M},\ \ pen(m) = (1+\theta)\frac{Tr({}^tP_n\pi_mP_n)}{n}\hat{\sigma}^2 \ . \tag{25}$$

For any $0<q\leqslant 1$ *such that* $2(q+1)<p$*, the penalized least-squares estimator* $\tilde{s}$ *satisfies*

$$\mathbb{E}[\|s-\tilde{s}\|_n^{2q}]^{1/q} \leqslant C\left(\inf_{m\in\mathcal{M}}\mathbb{E}[\|s-\hat{s}_m\|_n^2]+2\|s-\pi s\|_n^2\right)+\rho^2(P_n)\sigma^2\bar{R}_n(p,q,\theta)$$

where $C=C(q,\theta)$ *and* $C'=C'(p,q,\theta)$ *are positive constants,* $\bar{R}_n(p,q,\theta)$ *is equal to*

$$\frac{R_n(p,q,\theta)^{1/q}}{n}+C'\tau_p^{1/q}\kappa_n\left(\frac{\|s\|_n^2}{\rho^2(P_n)\sigma^2}+\tau_p\right)\left(\frac{\rho^{2\alpha_p}(P_n)}{Tr({}^tP_nP_n)^{\beta_p}}\right)^{1/q-2/p}$$

with $R_n(p,q,\theta)$ *defined as in Theorem 2.2,* $(\kappa_n)_{n\in\mathbb{N}}=(\kappa_n(p,q,\theta))_{n\in\mathbb{N}}$ *is a sequence of positive numbers that tends to* $\kappa=\kappa(p,q,\theta)>0$ *as* $Tr({}^tP_nP_n)/\rho^2(P_n)$ *increases toward infinity and*

$$\alpha_p=(p/2-1)\vee 1 \ \text{ and } \ \beta_p=(p/2-1)\wedge 1 \ .$$

Penalties given by (23) and (25) are random and allow to construct estimators $\tilde{s}$ when σ^2 is unknown. This approach leads to theoretical upperbounds for the risk of $\tilde{s}$. Note that we use some generic model V to construct $\hat{\sigma}^2$. This space is quite arbitrary and is pretty much limited to be an half-space of $\mathrm{Im}(P_n)$. The idea is that taking V as some "large" space can lead to a good approximation of the true s and, thus, $Y - \pi Y$ is not far from being centered and its normalized norm is of order σ^2. However, in practice, it is known that the estimator $\hat{\sigma}^2$ inclined to overestimate the true value of σ^2 as illustrated by Lemmas 8.4 and 8.5. Consequently, the penalty function tends to be larger and the procedure overpenalizes models with high dimension. To offset this phenomenon, a practical solution could be to choose some smaller θ when σ^2 is unknown than when it is known as we discuss in Section 6.

4 Application to additive models

In this section, we focus on the framework (4) given by an additive model. To describe the procedure to estimate the component s, we assume that the variance factor σ^2 is known but it can be easily generalized to the unknown factor case by considering the results of Section 3. We recall that $s \in \mathbb{L}_0^2([0,1],\nu)$, $t^j \in \mathbb{L}_0^2([0,1],\nu_j)$, $j = 1,\dots,K$, and we observe

$$Z_i = s_i + t_i + \sigma\varepsilon_i,\ i = 1,\dots,n\ , \tag{26}$$

where the random vector $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)'$ is such that $\mathbf{(H_{Gau})}$ or $\mathbf{(H_{Mom})}$ holds and the vectors $s = (s_1,\dots,s_n)'$ and $t = (t_1,\dots,t_n)'$ are defined in (7).

Let $\mathcal{S}_n$ be a linear subspace of $\mathbb{L}_0^2([0,1],\nu)$ and, for all $j \in \{1,\dots,K\}$, $\mathcal{S}_n^j$ be a linear subspace of $\mathbb{L}_0^2([0,1],\nu_j)$. We assume that these spaces have finite dimensions $D_n = \dim(\mathcal{S}_n)$ and $D_n^{(j)} = \dim(\mathcal{S}_n^j)$ such that

$$D_n + D_n^{(1)} + \cdots + D_n^{(K)} < n\ .$$

We consider an orthonormal basis $\{\phi_1,\dots,\phi_{D_n}\}$ (resp. $\{\psi_1^{(j)},\dots,\psi_{D_n^{(j)}}^{(j)}\}$) of $\mathcal{S}_n$ (resp. $\mathcal{S}_n^j$) equipped with the usual scalar product of $\mathbb{L}^2([0,1],\nu)$ (resp. of $\mathbb{L}^2([0,1],\nu_j)$). The linear spans $E, F^1,\dots,F^K \subset \mathbb{R}^n$ are defined by

$$E = \mathrm{Span}\left\{(\phi_i(x_1),\dots,\phi_i(x_n))',\ i = 1,\dots,D_n\right\}$$

and

$$F^j = \mathrm{Span}\left\{(\psi_i^{(j)}(y_1^j),\dots,\psi_i^{(j)}(y_n^j))',\ i = 1,\dots,D_n^{(j)}\right\},\ j = 1,\dots,K\ .$$

Let $\mathbf{1}_n = (1,\dots,1)' \in \mathbb{R}^n$, we also define

$$F = \mathbb{R}\mathbf{1}_n + F^1 + \cdots + F^K$$

where $\mathbb{R}\mathbf{1}_n$ is added to the F^j's in order to take into account the constant part μ of (4). Furthermore, note that the sum defining the space F does not need to be direct.

We are free to choose the functions ϕ_i's and ψ_i^j's. In the sequel, we assume that these functions are chosen in such a way that the mild assumption $E \cap F = \{0\}$ is fulfilled. Note that we do not assume that s belongs to E neither that t belongs to F. Let G be the space $(E+F)^\perp$, we obviously have $E \oplus F \oplus G = \mathbb{R}^n$ and we denote by P_n the projection onto E

along $F+G$. Moreover, we define π_E and π_{F+G} as the orthogonal projections onto E and $F+G$ respectively. Thus, we derive the following framework from (26),

$$Y = P_n Z = \bar{s} + \sigma P_n \varepsilon \tag{27}$$

where we have set

$$\begin{aligned} \bar{s} &= P_n s + P_n t \\ &= s + (P_n - I_n)s + P_n t \\ &= s + (P_n - I_n)(s - \pi_E s) + P_n(t - \pi_{F+G} t) = s + h \ . \end{aligned}$$

Let $\mathcal{F} = \{S_m,\ m \in \mathcal{M}\}$ be a finite collection of linear subspaces of E, we apply the procedure described in Section 2 to Y given by (27), that is, we choose an index $\hat{m} \in \mathcal{M}$ as a minimizer of (11) with a penalty function satisfying the hypotheses of Theorems 2.1 or 2.2 according to whether $\mathbf{(H_{Gau})}$ or $\mathbf{(H_{Mom})}$ holds. This way, we estimate s by $\tilde{s}$. From the triangular inequality, we derive that

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leqslant 2\mathbb{E}[\|\bar{s} - \tilde{s}\|_n^2] + 2\|h\|_n^2 \ .$$

As we discussed previously, under suitable assumptions on the complexity of the collection $\mathcal{F}$, we can assume that $\mathbf{(A_1)}$ and $\mathbf{(A_2)}$ are fulfilled. Let us suppose for the moment that $\mathbf{(A_3)}$ is satisfied for some $\rho > 1$. Note that, for any $m \in \mathcal{M}$, π_m is an orthogonal projection onto the image set of the oblique projection P_n. Consequently, we have $\mathrm{Tr}({}^tP_n\pi_m P_n) \geqslant \mathrm{rk}(\pi_m) = \dim(S_m)$ and Assumption $\mathbf{(A_3)}$ implies $\mathbf{(A_3')}$ with $c = 1/\rho^2$. Since, for all $m \in \mathcal{M}$,

$$\|\bar{s} - \pi_m \bar{s}\|_n \leqslant \|s - \pi_m s\|_n + \|h - \pi_m h\|_n \leqslant \|s - \pi_m s\|_n + \|h\|_n \ ,$$

we deduce from Theorems 2.1 or 2.2 that we can find, independently from s and n, two positive numbers C and C' such that

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leqslant C \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 \right\} + C' \left(\|h\|_n^2 + \frac{\rho^2\sigma^2}{n} R \right) \ . \tag{28}$$

To derive an interesting upperbound on the $\mathbb{L}^2$-risk of $\tilde{s}$, we need to control the remainder term. Because $\rho(\cdot)$ is a norm on $\mathbb{M}_n$, we dominate the norm of h by

$$\begin{aligned} \|h\|_n &\leqslant \rho(I_n - P_n)\|s - \pi_E s\|_n + \rho(P_n)\|t - \pi_{F+G} t\|_n \\ &\leqslant (1 + \rho(P_n))(\|s - \pi_E s\|_n + \|t - \pi_{F+G} t\|_n) \\ &\leqslant (1+\rho)(\|s - \pi_E s\|_n + \|t - \pi_{F+G} t\|_n) \ . \end{aligned}$$

Note that, for any $m \in \mathcal{M}$, $S_m \subset E$ and so, $\|s - \pi_E s\|_n \leqslant \|s - \pi_m s\|_n$. Thus, Inequality (28) leads to

$$\mathbb{E}[\|s-\tilde{s}\|_n^2] \leqslant C(1+\rho)^2 \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 \right\} + C'(1+\rho)^2 \left(\|t - \pi_{F+G} t\|_n^2 + \frac{\sigma^2}{n} R \right) \ . \tag{29}$$

The space $F+G$ has to be seen as a large approximation space. So, under a reasonable assumption on the regularity of the component t, the quantity $\|t - \pi_{F+G} t\|_n^2$ could be regarded as being neglectable. It mainly remains to understand the order of the multiplicative factor $(1+\rho)^2$.

Thus, we now discuss about the norm $\rho(P_n)$ and the assumption $(\mathbf{A}_3)$. This quantity depends on the design points $(x_i, y_i^1, \dots, y_i^K) \in [0,1]^{K+1}$ and on how we construct the spaces E and F, *i.e.* on the choice of the basis functions ϕ_i and $\psi_i^{(j)}$. Hereafter, the design points $(x_i, y_i^1, \dots, y_i^K)$ will be assumed to be known independent realizations of a random variable on $[0,1]^{K+1}$ with distribution $\nu \otimes \nu_1 \otimes \cdots \otimes \nu_K$. We also assume that these points are independent of the noise ε and we proceed conditionally to them. To discuss about the probability for $(\mathbf{A}_3)$ to occur, we introduce some notations. We denote by D_n' the integer

$$D_n' = 1 + D_n^{(1)} + \cdots + D_n^{(K)}$$

and we have $\dim(F) \leqslant D_n'$. Let A be a $p \times p$ real matrix, we define

$$r_p(A) = \sup \left\{ \sum_{i=1}^{p} \sum_{j=1}^{p} |a_i a_j| \times |A_{ij}| \ : \sum_{i=1}^{p} a_i^2 \leqslant 1 \right\} .$$

Moreover, we define the matrices $V(\phi)$ and $B(\phi)$ by

$$V_{ij}(\phi) = \sqrt{\int_0^1 \phi_i(x)^2 \phi_j(x)^2 \nu(dx)} \quad \text{and} \quad B_{ij}(\phi) = \sup_{x \in [0,1]} |\phi_i(x)\phi_j(x)| ,$$

for any $1 \leqslant i, j \leqslant D_n$. Finally, we introduce the quantities

$$L_\phi = \max \left\{ r_{D_n}^2(V(\phi)),\ r_{D_n}(B(\phi)) \right\} \quad \text{and} \quad b_\phi = \max_{i=1,\dots,D_n} \sup_{x \in [0,1]} |\phi_i(x)|$$

and

$$L_n = \max \left\{ L_\phi,\ D_n D_n',\ b_\phi \sqrt{n D_n D_n'} \right\} .$$

Proposition 4.1. *Consider the matrix P_n defined in (27). We assume that the design points are independent realizations of a random variable on $[0,1]^{K+1}$ with distribution $\nu \otimes \nu_1 \otimes \cdots \otimes \nu_K$ such that we have $E \cap F = \{0\}$ and $\dim(E) = D_n$ almost surely. If the basis $\{\phi_1, \dots, \phi_{D_n}\}$ is such that*

$$\forall 1 \leqslant i \leqslant D_n, \ \int_0^1 \phi_i(x) \nu(dx) = 0 \tag{30}$$

then, there exists some universal constant $C > 0$ such that, for any $\rho > 1$,

$$\mathbb{P}\left(\rho(P_n) > \rho\right) \leqslant 4 D_n (D_n + D_n') \exp\left(-\frac{Cn}{L_n} (1 - \rho^{-1})^2 \right) .$$

As a consequence of Proposition 4.1, we see that $(\mathbf{A}_3)$ is fulfilled with a large probability since we choose basis functions ϕ_i in such a way to keep L_n small in front of n. It will be so for localized bases (piecewise polynomials, orthonormal wavelets, ...) with L_n of order of $n^{1-\omega}$, for some $\omega \in (0,1)$, once we consider D_n and D_n' of order of $n^{\frac{1}{3} - \frac{3\omega}{2}}$ (this is a direct consequence of Lemma 1 in [8]). This limitation, mainly due to the generality of the proposition, could seem restrictive from a practical point of view. However the statistician can explicitly compute $\rho(P_n)$ with the data. Thus, it is possible to adjust D_n and D_n' in order to keep $\rho(P_n)$ small in practice. Moreover, we will see in Section 6 that, for our choices of ϕ_i and ψ_i^j, we can easily consider D_n and D_n' of order of $\sqrt{n}$ as we keep $\rho(P_n)$ small (concrete values are given in the simulation study).

5 Convergence rates

The previous sections have introduced various upperbounds on the $\mathbb{L}^2$-risk of the penalized least-squares estimators $\tilde{s}$. Each of them is connected to the minimal risk of the estimators among a collection $\{\hat{s}_m, m \in \mathcal{M}\}$. One of the main advantages of such inequalities is that it allows us to derive uniform convergence rates with respect to many well known classes of smoothness (see [7]). In this section, we give such results over Hölderian balls for the estimation of a component in an additive framework. To this end, for any $\alpha > 0$ and $R > 0$, we introduce the space $\mathcal{H}_\alpha(R)$ of the α-Hölderian functions with constant $R > 0$ on $[0,1]$,

$$\mathcal{H}_\alpha(R) = \{f : [0,1] \to \mathbb{R} \ : \ \forall x, y \in [0,1], \ |f(x) - f(y)| \leqslant R|x-y|^\alpha\} \ .$$

In order to derive such convergence rates, we need a collection of models $\mathcal{F}$ with good approximation properties for the functions of $\mathcal{H}_\alpha(R)$. We denote by P_n^{BM} any oblique projector defined as in the previous section and based on spaces $\mathcal{S}_n$ and $\mathcal{S}_n^j$ that are constructed as one of the examples given in Section 2 of [9]. In particular, such a construction allows us to deal with approximation spaces $\mathcal{S}_n$ and $\mathcal{S}_n^j$ that can be considered as spaces of piecewise polynomials, spaces of orthogonal wavelet expansions or spaces of dyadic splines on $[0,1]$. We consider the dimensions $D_n = \dim(\mathcal{S}_n)$ and, for any $j \in \{1, \ldots, K\}$, $D_n^{(j)} = \dim(\mathcal{S}_n^j) = D_n/K$. Finally, we take a collection of models $\mathcal{F}^{BM}$ that contains subspaces of $E = \mathrm{Im}(P_n^{BM})$ as Baraud did in Section 2.2 of [5].

Proposition 5.1. *Consider the framework (4) and assume that* $(\mathbf{H_{Gau}})$ *or* $(\mathbf{H_{Mom}})$ *holds with $p > 6$. We define Y in (27) with P_n^{BM}. Let $\eta > 0$ and $\tilde{s}$ be the estimator selected by the procedure (11) applied to the collection of models $\mathcal{F}^{BM}$ with the penalty*

$$pen(m) = (1+\eta)\frac{Tr({}^tP_n^{BM}\pi_m P_n^{BM})}{n}\sigma^2 \ .$$

Suppose that $(\mathbf{A}_3)$ *is fulfilled, we define*

$$\zeta_n = \frac{1}{2}\left(\frac{\log n}{\log D_n} - 1\right) > 0 \ .$$

For any $\alpha > \zeta_n$ and $R > 0$, the penalized least-squares estimator $\tilde{s}$ satisfies

$$\sup_{(s,t^1,\ldots,t^K)\in\mathcal{H}_\alpha(R)^{K+1}} \mathbb{E}_{\varepsilon,d}\left[\|s - \tilde{s}\|_n^2\right] \leqslant C_\alpha n^{-2\alpha/(2\alpha+1)} \tag{31}$$

where $\mathbb{E}_{\varepsilon,d}$ is the expectation on ε and on the random design points and $C_\alpha > 1$ only depends on α, ρ, σ^2, K, L, θ and p (under $(\mathbf{H_{Mom}})$ *only).*

Note that the supremum is taken over Hölderian balls for all the components of the regression function, *i.e.* the regression function is itself supposed to belong to an Hölderian space. As we mention in the introduction, Stone [37] has proved that the rate of convergence given by (31) is optimal in the minimax sense.

6 Simulation study

In this section, we study simulations based on the framework given by (4) with $K+1$ components $s, t^1, \ldots, t^K$ and Gaussian errors. First, we introduce the spaces $\mathcal{S}_n$ and $\mathcal{S}_n^j$, $j \in \{1, \ldots, K\}$, and the collections of models that we handle. Next, we illustrate the performances of the estimators in practice by several examples.

6.1 Preliminaries

To perform the simulation study, we consider two collections of models. In both cases, we deal with the same spaces $\mathcal{S}_n$ and $\mathcal{S}_n^j$ defined as follows. Let φ be the Haar wavelet's mother function,

$$\forall x \in \mathbb{R},\ \varphi(x) = \begin{cases} 1 & \text{if } 0 \leqslant x < 1/2\ , \\ -1 & \text{if } 1/2 \leqslant x < 1\ , \\ 0 & \text{otherwise.} \end{cases}$$

For any $i \in \mathbb{N}$ and $j \in \{0, \dots, 2^i - 1\}$, we introduce the functions

$$\varphi_{i,j}(x) = 2^{i/2}\varphi(2^i x - j),\ x \in \mathbb{R}\ .$$

It is clear that these functions are orthonormal in $\mathbb{L}_0^2([0,1], dx)$ for the usual scalar product. Let d_n be some positive integer, we consider the space $\mathcal{S}_n \subset \mathbb{L}_0^2([0,1], dx)$ generated by the functions $\varphi_{i,j}$ such that $0 \leqslant i \leqslant d_n$ and $0 \leqslant j < 2^i$. The dimension of this space is $\dim(\mathcal{S}_n) = D_n = 2^{d_n+1} - 1$. In the sequel, we denote by Π_n the set of all the allowed pairs (i,j),

$$\Pi_n = \left\{(i,j) \in \mathbb{N}^2 \text{ such that } 0 \leqslant i \leqslant d_n,\ 0 \leqslant j < 2^i\right\}\ .$$

Moreover, for any $k \in \{1, \dots, D_n\}$ such that $k = 2^i + j$ with $(i,j) \in \Pi_n$, we denote $\phi_k = \varphi_{i,j}$.

Let d_n' be an other positive integer, the spaces $\mathcal{S}_n^j \subset \mathbb{L}_0^2([0,1], dy^j)$ are all supposed to be generated by the functions defined on $[0,1]$ by

$$\psi_{2i}(y) = \psi_{2i}^{(j)}(y) = \sin(i\pi y) \quad \text{and} \quad \psi_{2i-1}(y) = \psi_{2i-1}^{(j)}(y) = \cos(i\pi y)$$

for any $i \in \{1, \dots, d_n'\}$ and $j \in \{1, \dots, K\}$. Thus, we have $\dim(\mathcal{S}_n^j) = D_n^{(j)} = 2d_n'$ and $D_n' = 2Kd_n' + 1$.

As previously, we define P_n as the oblique projector onto E along $F + (E+F)^\perp$. The image set $E = \mathrm{Im}(P_n)$ is generated by the vectors

$$\varphi_{i,j} = (\varphi_{i,j}(x_1), \dots, \varphi_{i,j}(x_n))' \in \mathbb{R}^n,\ (i,j) \in \Pi_n\ .$$

Let m be a subset of Π_n, the model S_m is defined as the linear subspace of E generated by the vectors $\varphi_{i,j}$ with $(i,j) \in m$.

In the following simulations, we always take D_n and D_n' close to $4\sqrt{n}$, *i.e.*

$$d_n = \left\lfloor \frac{\log(2\sqrt{n} + 1/2)}{\log(2)} \right\rfloor \quad \text{and} \quad d_n' = \left\lfloor \frac{4\sqrt{n} - 1}{2K} \right\rfloor$$

where, for any $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer not greater than x. For such choices, basic computations lead to L_n of order of $n^{5/4}$ in Proposition 4.1. As a consequence, this proposition does not ensure that $(\mathbf{A}_3)$ is fulfilled with a large probability. However, $\rho(P_n)$ remains small in practice as we will see and it allows us to deal with larger collections of models.

6.2 Collections of models

The first collection of models is the smaller one because the models are nested. Let us introduce the index subsets, for any $i \in \{0, \dots, d_n\}$,

$$m_i = \left\{(i,j),\ 0 \leqslant j < 2^i\right\} \subset \Pi_n\ .$$

Thus, we define $\mathcal{F}^N$ as

$$\mathcal{F}^N = \left\{ S_m \text{ such that } \exists k \in \{0, \dots, d_n\},\ m = \bigcup_{i=0}^{k} m_i \right\} .$$

This collection has a small complexity since, for any $d \in \mathbb{N}$, $N_d \leqslant 1$. According to Corollary 2.1, we can consider the penalty function given by

$$\mathrm{pen}_N(m) = (1 + C)\frac{\mathrm{Tr}({}^tP_n \pi_m P_n)}{n}\sigma^2 \tag{32}$$

for some $C > 0$. In order to compute the selected estimator $\tilde{s}$, we simply compute $\hat{s}_m$ in each model of $\mathcal{F}^N$ and we take the one that minimizes the penalized least-squares criterion.

The second collection of models is larger than $\mathcal{F}^N$. Indeed, we allow m to be any subset of Π_n and we introduce

$$\mathcal{F}^C = \{ S_m \text{ such that } m \subset \Pi_n \} .$$

The complexity of this collection is large because, for any $d \in \mathbb{N}$,

$$N_d = \binom{D_n}{d} = \frac{D_n!}{d!(D_n - d)!} \leqslant \left(\frac{eD_n}{d}\right)^d .$$

So, we have $\log N_d \leqslant d(1 + \log D_n)$ and, according to Corollary 2.1, we take a penalty function as

$$\mathrm{pen}_C(m) = (1 + C + \log D_n)\frac{\mathrm{Tr}({}^tP_n \pi_m P_n)}{n}\sigma^2 \tag{33}$$

for some $C > 0$. The large number of models in $\mathcal{F}^C$ leads to difficulties for computing the estimator $\tilde{s}$. Instead of exploring all the models among $\mathcal{F}^C$, we break the penalized criterion down with respect to an orthonormal basis $\phi_1, \dots, \phi_{D_n}$ of E and we get

$$\begin{aligned} & \left\| Y - \sum_{i=1}^{D_n} \langle Y, \phi_i \rangle_n \phi_i \right\|_n^2 + (1 + C + \log D_n)\frac{\mathrm{Tr}({}^tP_n \pi_E P_n)}{n}\sigma^2 \\ = \ & \|Y\|_n^2 - \sum_{i=1}^{D_n} \left[\langle Y, \phi_i \rangle_n^2 - (1 + C + \log D_n) \| {}^tP_n \phi_i \|_n^2 \sigma^2 \right] . \end{aligned}$$

In order to minimize the penalized least-squares criterion, we only need to keep the coefficients $\langle Y, \phi_i \rangle_n$ that are such that

$$\langle Y, \phi_i \rangle_n^2 \geqslant (1 + C + \log D_n) \| {}^tP_n \phi_i \|_n^2 \sigma^2 .$$

This threshold procedure allows us to compute the estimator $\tilde{s}$ in reasonable time.

In accordance with the results of Section 3, in the case of unknown variance, we substitute $\hat{\sigma}^2$ for σ^2 in the penalties (32) and (33).

6.3 Numerical simulations

We now illustrate our results and the performances of our estimation procedure by applying it to simulated data

$$Z_i = s(x_i) + \sum_{j=1}^{K} t^j(y_i^j) + \sigma \varepsilon_i,\ i = 1, \dots, n ,$$

where $K \geqslant 1$ is an integer that will vary from an experiment to an other, the design points $(x_i, y_i^1, \ldots, y_i^K)'$ are known independent realizations of an uniform random variable on $[0,1]^{K+1}$ and the errors ε_i are i.i.d. standard Gaussian random variables. We handle this framework with known or unknown variance factor $\sigma^2 = 1$ according to the cases and we consider a design of size $n = 512$. The unknown components $s, t^1, \ldots, t^K$ are either chosen among the following ones, or set to zero in the last subsection,

$$f_1(x) = \sin\left(4\pi\left(x \wedge \frac{1}{2}\right)\right) \qquad f_2(x) = \cos\left(2\pi\left(x - \frac{1}{4}\right)^2\right) - C_2 \quad f_3(x) = x + 2\exp(-16x^2) - C_3$$

$$f_4(x) = \sin(2x) + 2\exp(-16x^2) - C_4 \quad f_5(x) = \frac{1 - \exp(-10(x - 1/2))}{1 + \exp(-10(x - 1/2))} \qquad f_6(x) = 6x(1 - x) - 1$$

where the constants C_2, C_3 and C_4 are such that $f_i \in \mathbb{L}_0^2([0,1], dx)$ for any $i \in \{1, \ldots, 6\}$.

The first step of the procedure consists in computing the oblique projector P_n and taking the data $Y = P_n Z$. Figure 1 gives an example by representing the signal s, the data Z and the projected data Y for $K = 6$, $s = f_1$ and $t^j = f_j$, $j \in \{1, \ldots, 6\}$. In particular, for this example, we have $\rho^2(P_n) = 1.22$. We see that we actually get reasonable value of $\rho^2(P_n)$ with our particular choices for D_n and D_n'.

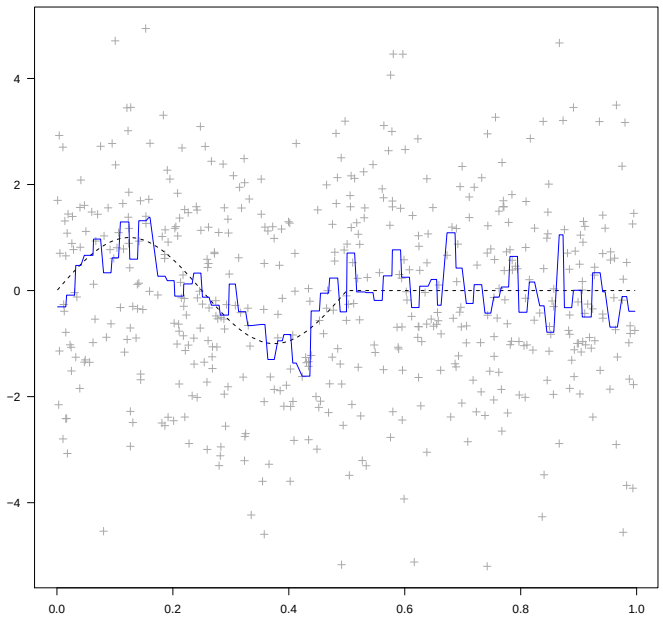

Figure 1: Plot in (x, z) of the signal s (dashed line), the data Z (dots) and the projected data Y (plain line).

In order to estimate the component s, we choose $\hat{m}$ by the procedure (11) with penalty

function given by (32) or (33) according to the cases. The first simulations deal with the collection $\mathcal{F}^N$ of nested models. Figure 2 represents the true s and the estimator $\tilde{s}$ for $K = 6$ parasitic components given by $t^j = f_j,\ j \in \{1, \dots, 6\}$ and $s = f_1$ or $s = f_5$. The penalty function (32) has been used with a constant $C = 1.5$.

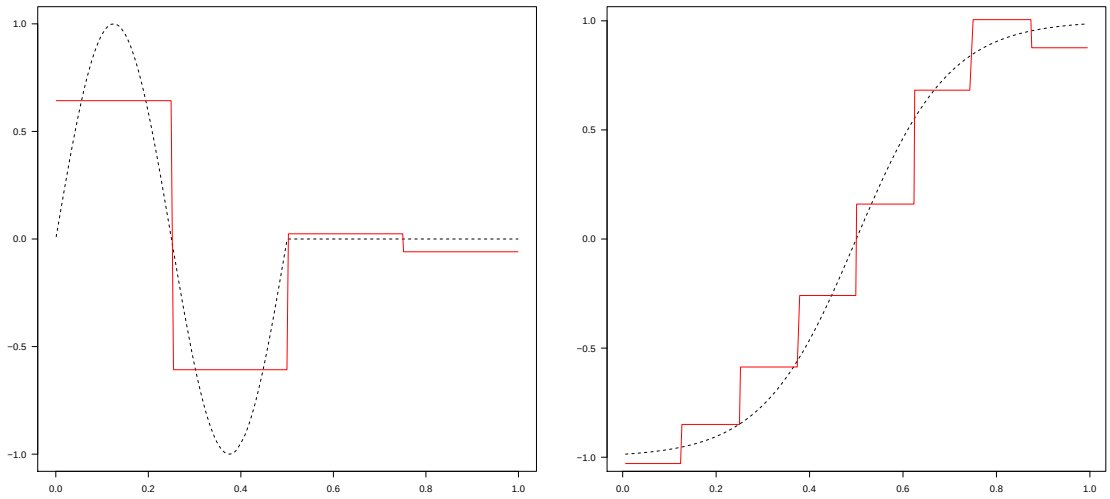

Figure 2: Estimation of s (dashed) by $\tilde{s}$ (plain) with $\mathcal{F}^N$, $K = 6$ and $t^j = f_j$, $j \in \{1, \dots, 6\}$, for $s = f_1$ (left, $\rho(P_n) = 1.24$) and for $s = f_5$ (right, $\rho(P_n) = 1.25$).

The second set of simulations is related to the large collection $\mathcal{F}^C$ and to the penalty function (33) with $C = 4.5$. Figure 3 illustrates the estimation of $s = f_1$ and $s = f_2$ with $K = 6$ parasitic components $t^j = f_j,\ j \in \{1, \dots, 6\}$.

In both cases, we see that the estimation procedure behaves well and that the norms $\rho(P_n)$ are close to one in spite of the presence of the parasitic components. Moreover, note that the collection $\mathcal{F}^C$ allows to get estimators that are sharper because they detect constant parts of s. This advantage leads to a better bias term in the quadratic risk decomposition at the price of the logarithmic term in the penalty (33).

6.4 Ratio estimation

In Section 4, we discussed about assumptions that ensure a small remainder term in Inequality (29). This result corresponds to some oracle type inequality for our estimation procedure of a component in an additive framework. We want to evaluate how far $\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right]$ is from the oracle risk. Thus, we estimate the ratio

$$r_K(\tilde{s}) = \frac{\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right]}{\inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \dfrac{\mathrm{Tr}({}^tP_n \pi_m P_n)}{n} \sigma^2 \right\}}$$

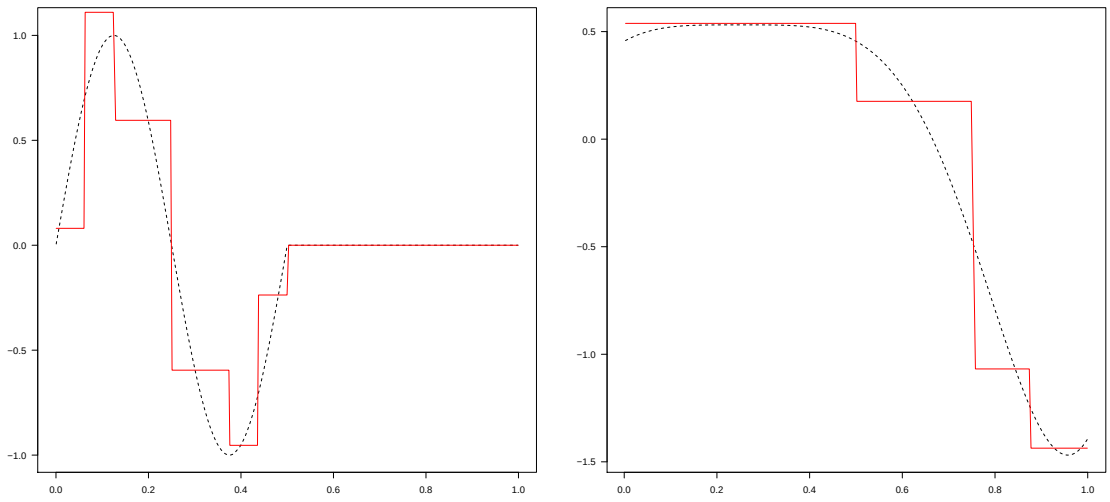

Figure 3: Estimation of s (dashed) by $\tilde{s}$ (plain) with $\mathcal{F}^C$, $K = 6$ and $t^j = f_j$, $j \in \{1, \dots, 6\}$, for $s = f_1$ (left, $\rho(P_n) = 1.23$) and for $s = f_2$ (right, $\rho(P_n) = 1.27$).

by repeating 500 times each experiment for various values of K and C. For each set of simulations, the parasitic components are taken such that $t^j = f_j$, $j \in \{1, \dots, K\}$, the values of $\rho(P_n)$ are given and the variance σ^2 is either assumed to be known or not.

Table 1 (resp. Table 2) gives the values of $r_K(\tilde{s})$ obtained for $s = f_1$ (resp. $s = f_5$) with the collection $\mathcal{F}^N$ and the penalty (32). We clearly see that taking C close to zero or too large is not a good thing for the procedure. In our examples, $C = 1.5$ give good results and we get reasonable values of $r_K(\tilde{s})$ for other choices of C between 1 and 3 for known or unknown variance. As expected, we also note that the values of $\rho(P_n)$ and $r_K(\tilde{s})$ tend to increase when K goes up but remain acceptable for $K \in \{1, \dots, 6\}$.

In the same way, we estimate the ratio $r_K(\tilde{s})$ for $s = f_1$ and $s = f_2$ with the collection $\mathcal{F}^C$ and the penalty (33). The results are given in Table 3 and Table 4. We obtain reasonable values of $r_K(\tilde{s})$ for choices of C larger than what we took in the nested case. This phenomenon is related to what we mentioned at the end of Section 2. Indeed, for large collection of models, we need to overpenalize in order to keep the remainder term small enough. Moreover, because $\hat{\sigma}^2$ tends to overestimate σ^2 (see Section 3), we see that we can consider smaller values for C when the variance is unknown than when it is known for obtaining equivalent results.

6.5 Parasitic components equal to zero

We are now interested in the particular case of parasitic components t^j equal to zero in (4), i.e. data are given by

$$Z_i = s(x_i) + \sigma\varepsilon_i, \ i = 1, \dots, n \ .$$

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K = 1$,	2.41	1.36	1.15	1.13	1.11	1.10	1.09	1.08	1.08	1.08	1.08
$\rho(P_n) = 1.23$	1.46	1.29	1.19	1.14	1.10	1.09	1.09	1.09	1.08	1.08	1.08
$K = 2$,	2.47	1.37	1.16	1.14	1.13	1.12	1.11	1.09	1.09	1.09	1.09
$\rho(P_n) = 1.23$	1.55	1.26	1.18	1.14	1.12	1.12	1.11	1.10	1.09	1.09	1.09
$K = 3$,	2.48	1.39	1.15	1.13	1.12	1.10	1.09	1.08	1.08	1.08	1.08
$\rho(P_n) = 1.28$	2.34	1.26	1.16	1.13	1.11	1.10	1.09	1.09	1.08	1.08	1.08
$K = 4$,	2.65	1.41	1.17	1.14	1.13	1.11	1.09	1.08	1.08	1.08	1.08
$\rho(P_n) = 1.25$	1.46	1.27	1.16	1.13	1.11	1.10	1.09	1.09	1.08	1.08	1.08
$K = 5$,	2.97	1.62	1.27	1.19	1.15	1.12	1.10	1.09	1.08	1.07	1.07
$\rho(P_n) = 1.29$	1.63	1.38	1.26	1.19	1.13	1.11	1.09	1.08	1.08	1.08	1.07
$K = 6$,	3.14	1.77	1.29	1.21	1.17	1.13	1.12	1.10	1.10	1.09	1.09
$\rho(P_n) = 1.27$	1.66	1.40	1.26	1.18	1.14	1.13	1.11	1.11	1.10	1.10	1.09

Table 1: Ratio $r_K(\tilde{s})$ for the estimation of $s = f_1$ with $\mathcal{F}^N$. Each pair of lines corresponds to a value of K with the known σ^2 case on the first line and unknown σ^2 case on the second one.

If we know that these K components are zero and if we deal with the collection $\mathcal{F}^N$ and a known variance σ^2, we can consider the classical model selection procedure given by

$$\hat{m}_0 \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \|Z - \pi_m Z\|_n^2 + C \frac{\dim(S_m)}{n} \sigma^2 \right\} . \tag{34}$$

Then, we can define the estimator $\tilde{s}_0 = \pi_{\hat{m}_0} Z$. This procedure is well known and we refer to [27] for more details. If we do not know that the K parasitic components are null, we can use our procedure to estimate s by $\tilde{s}$. In order to compare the performances of $\tilde{s}$ and $\tilde{s}_0$ with respect to the number K of zero parasitic components, we estimate the ratio

$$r_K(\tilde{s}, \tilde{s}_0) = \frac{\mathbb{E}[\|s - \tilde{s}\|_n^2]}{\mathbb{E}[\|s - \tilde{s}_0\|_n^2]}$$

for various values of K and C by repeating 500 times each experiment.

The obtained results are given in Tables 5 and 6 for $s = f_1$ and $s = f_5$ respectively. Obviously, the ratio $r_K(\tilde{s}, \tilde{s}_0)$ is always larger than one because the procedure (34) makes good use of its knowledge about nullity of the t^j. Nevertheless, we see that our procedure performs nearly as well as (34) even for a large number of zero components. Indeed, for $K \in \{1, \dots, 9\}$, do not assuming that we know that the t^j are zero only implies a loss between 1% and 10% for the risk. Such a loss remains acceptable in practice and allows us to consider more general framework for estimating s.

7 Proofs

In the proofs, we repeatedly use the following elementary inequality that holds for any $\alpha > 0$ and $x, y \in \mathbb{R}$,

$$2|xy| \leqslant \alpha x^2 + \alpha^{-1} y^2 . \tag{35}$$

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K = 1$,	4.08	1.52	1.22	1.20	1.27	1.35	1.45	1.56	1.64	1.70	1.79
$\rho(P_n) = 1.28$	3.44	1.58	1.36	1.26	1.30	1.37	1.45	1.55	1.64	1.72	1.81
$K = 2$,	4.07	1.66	1.28	1.26	1.32	1.40	1.49	1.57	1.66	1.74	1.82
$\rho(P_n) = 1.23$	2.29	1.69	1.36	1.32	1.36	1.44	1.53	1.60	1.65	1.73	1.82
$K = 3$,	4.17	1.65	1.36	1.34	1.42	1.50	1.60	1.67	1.77	1.89	2.01
$\rho(P_n) = 1.25$	2.24	1.70	1.41	1.41	1.48	1.55	1.61	1.71	1.80	1.92	2.01
$K = 4$,	4.42	1.88	1.43	1.34	1.36	1.45	1.53	1.61	1.69	1.77	1.86
$\rho(P_n) = 1.26$	3.80	1.75	1.51	1.42	1.44	1.50	1.56	1.66	1.75	1.84	1.93
$K = 5$,	4.57	1.82	1.43	1.37	1.39	1.46	1.53	1.60	1.67	1.76	1.83
$\rho(P_n) = 1.26$	2.33	1.77	1.51	1.43	1.44	1.50	1.54	1.64	1.74	1.82	1.89
$K = 6$,	4.98	2.08	1.59	1.47	1.45	1.49	1.57	1.66	1.77	1.86	1.96
$\rho(P_n) = 1.27$	2.57	1.91	1.62	1.52	1.54	1.57	1.65	1.73	1.84	1.93	2.02

Table 2: Ratio $r_K(\tilde{s})$ for the estimation of $s = f_5$ with $\mathcal{F}^N$. Each pair of lines corresponds to a value of K with the known σ^2 case on the first line and unknown σ^2 case on the second one.

7.1 Proofs of Theorems 2.1 and 2.2

7.1.1 Proof of Theorem 2.1

By definition of γ_n, for any $t \in \mathbb{R}^n$, we can write

$$\|s - t\|_n^2 = \gamma_n(t) + 2\sigma\langle t - Y, P_n\varepsilon\rangle_n + \sigma^2\|P_n\varepsilon\|_n^2 \ .$$

Let $m \in \mathcal{M}$, since $\hat{s}_m = s_m + \sigma\pi_m P_n\varepsilon$, this identity and (12) lead to

$$\begin{aligned}
\|s - \tilde{s}\|_n^2 &= \|s - s_m\|_n^2 + \gamma_n(\tilde{s}) - \gamma_n(s_m) + 2\sigma\langle\tilde{s} - s_m, P_n\varepsilon\rangle_n \\
&= \|s - s_m\|_n^2 + \gamma_n(\tilde{s}) - \gamma_n(\hat{s}_m) - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \\
&\quad -2\sigma\langle s - \tilde{s}, P_n\varepsilon\rangle_n + 2\sigma\langle s - s_m, P_n\varepsilon\rangle_n \\
&\leqslant \|s - s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + 2\sigma^2\|\pi_{\hat{m}} P_n\varepsilon\|_n^2 \\
&\quad -2\sigma\langle s - s_{\hat{m}}, P_n\varepsilon\rangle_n + 2\sigma\langle s - s_m, P_n\varepsilon\rangle_n - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \ .
\end{aligned} \tag{36}$$

Consider an arbitrary $a_m \in S_m^\perp$ such that $\|a_m\|_n = 1$, we define

$$u_m = \begin{cases} (s - s_m)/\|s - s_m\|_n & \text{if } s \neq \pi_m s \\ a_m & \text{otherwise .} \end{cases} \tag{37}$$

Thus, (36) gives

$$\begin{aligned}
\|s - \tilde{s}\|_n^2 &\leqslant \|s - s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + 2\sigma^2\|\pi_{\hat{m}} P_n\varepsilon\|_n^2 \\
&\quad +2\sigma\|s - s_{\hat{m}}\|_n|\langle u_{\hat{m}}, P_n\varepsilon\rangle_n| + 2\sigma\langle s - s_m, P_n\varepsilon\rangle_n - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \ .
\end{aligned} \tag{38}$$

Take $\alpha \in (0, 1)$ that we specify later and we use the inequality (35),

$$\begin{aligned}
(1 - \alpha)\|s - \tilde{s}\|_n^2 &\leqslant \|s - s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \\
&\quad +(2 - \alpha)\sigma^2\|\pi_{\hat{m}} P_n\varepsilon\|_n^2 + \alpha^{-1}\sigma^2\langle u_{\hat{m}}, P_n\varepsilon\rangle_n^2 \\
&\quad +2\sigma\langle s - s_m, P_n\varepsilon\rangle_n - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \ .
\end{aligned} \tag{39}$$

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K = 1$,	1.54	1.49	1.44	1.40	1.36	1.33	1.31	1.30	1.28	1.27	1.25
$\rho(P_n) = 1.27$	1.50	1.44	1.39	1.35	1.32	1.30	1.28	1.26	1.25	1.24	1.23
$K = 2$,	1.60	1.53	1.48	1.45	1.40	1.37	1.34	1.32	1.29	1.28	1.26
$\rho(P_n) = 1.25$	1.54	1.48	1.42	1.38	1.35	1.32	1.29	1.28	1.27	1.25	1.24
$K = 3$,	1.56	1.50	1.46	1.42	1.38	1.35	1.32	1.30	1.28	1.27	1.26
$\rho(P_n) = 1.25$	1.51	1.45	1.41	1.37	1.34	1.31	1.29	1.27	1.25	1.24	1.23
$K = 4$,	1.61	1.54	1.48	1.42	1.39	1.36	1.34	1.31	1.29	1.28	1.27
$\rho(P_n) = 1.25$	1.51	1.44	1.40	1.36	1.32	1.31	1.28	1.27	1.26	1.25	1.24
$K = 5$,	1.68	1.61	1.54	1.48	1.44	1.41	1.37	1.34	1.32	1.30	1.28
$\rho(P_n) = 1.25$	1.56	1.49	1.43	1.39	1.36	1.31	1.29	1.27	1.27	1.26	1.25
$K = 6$,	1.78	1.70	1.63	1.57	1.53	1.48	1.44	1.42	1.39	1.35	1.34
$\rho(P_n) = 1.24$	1.61	1.55	1.48	1.44	1.40	1.37	1.34	1.32	1.30	1.28	1.28

Table 3: Ratio $r_K(\tilde{s})$ for the estimation of $s = f_1$ with $\mathcal{F}^C$. Each pair of lines corresponds to a value of K with the known σ^2 case on the first line and unknown σ^2 case on the second one.

We choose $\alpha = 1/(1+\theta) \in (0,1)$ but for legibility we keep using the notation α. Let us now introduce two functions $p_1, p_2 : \mathcal{M} \to \mathbb{R}_+$ that will be specified later to satisfy, for all $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geqslant (2-\alpha)p_1(m) + \alpha^{-1}p_2(m) \ . \tag{40}$$

We use this bound in (39) to obtain

$$\begin{aligned}
(1-\alpha)\|s-\tilde{s}\|_n^2 \ \leqslant \ & \|s-s_m\|_n^2 + \mathrm{pen}(m) + (2-\alpha)\left(\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 - p_1(\hat{m})\right) \\
& + \alpha^{-1}\left(\sigma^2\langle u_{\hat{m}}, P_n\varepsilon\rangle_n^2 - p_2(\hat{m})\right) + 2\sigma\langle s-s_m, P_n\varepsilon\rangle_n \\
& - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \\
\leqslant \ & \|s-s_m\|_n^2 + \mathrm{pen}(m) + 2\sigma\langle s-s_m, P_n\varepsilon\rangle_n - \sigma^2\|\pi_m P_n\varepsilon\|_n^2 \\
& + (2-\alpha)\sup_{m'\in\mathcal{M}}\left(\sigma^2\|\pi_{m'}P_n\varepsilon\|_n^2 - p_1(m')\right)_+ \\
& + \alpha^{-1}\sup_{m'\in\mathcal{M}}\left(\sigma^2\langle u_{m'}, P_n\varepsilon\rangle_n^2 - p_2(m')\right)_+ \ .
\end{aligned}$$

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K = 1$,	2.01	1.92	1.86	1.80	1.76	1.74	1.70	1.70	1.68	1.67	1.68
$\rho(P_n) = 1.28$	2.03	1.93	1.87	1.81	1.77	1.72	1.68	1.65	1.65	1.66	1.67
$K = 2$,	2.02	1.93	1.85	1.79	1.75	1.71	1.68	1.66	1.66	1.66	1.66
$\rho(P_n) = 1.22$	1.95	1.88	1.82	1.78	1.75	1.71	1.68	1.67	1.65	1.64	1.64
$K = 3$,	2.04	1.93	1.86	1.81	1.76	1.71	1.68	1.64	1.62	1.62	1.62
$\rho(P_n) = 1.26$	1.96	1.87	1.80	1.74	1.68	1.66	1.63	1.63	1.61	1.62	1.62
$K = 4$,	2.12	2.00	1.90	1.81	1.73	1.67	1.64	1.62	1.60	1.61	1.60
$\rho(P_n) = 1.25$	1.99	1.90	1.80	1.73	1.68	1.65	1.62	1.60	1.60	1.60	1.60
$K = 5$,	2.47	2.34	2.23	2.17	2.10	2.05	1.99	1.95	1.91	1.88	1.86
$\rho(P_n) = 1.24$	2.30	2.20	2.11	2.03	1.97	1.92	1.88	1.83	1.82	1.80	1.80
$K = 6$,	2.45	2.32	2.21	2.11	2.03	1.99	1.95	1.91	1.89	1.86	1.84
$\rho(P_n) = 1.26$	2.17	2.06	1.99	1.94	1.89	1.85	1.84	1.80	1.79	1.79	1.75

Table 4: Ratio $r_K(\tilde{s})$ for the estimation of $s = f_2$ with $\mathcal{F}^C$. Each pair of lines corresponds to a value of K with the known σ^2 case on the first line and unknown σ^2 case on the second one.

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K = 1$	1.11	1.11	1.09	1.06	1.04	1.03	1.03	1.02	1.01	1.02	1.02
$K = 2$	1.12	1.08	1.08	1.06	1.04	1.03	1.02	1.01	1.01	1.01	1.01
$K = 3$	1.13	1.09	1.07	1.07	1.05	1.03	1.01	1.01	1.02	1.02	1.02
$K = 4$	1.08	1.08	1.06	1.05	1.04	1.02	1.02	1.01	1.01	1.01	1.01
$K = 5$	1.10	1.05	1.06	1.06	1.03	1.02	1.02	1.01	1.01	1.01	1.01
$K = 6$	1.08	1.07	1.06	1.05	1.03	1.02	1.01	1.01	1.01	1.01	1.01
$K = 7$	1.11	1.09	1.08	1.05	1.03	1.02	1.01	1.01	1.01	1.01	1.01
$K = 8$	1.09	1.06	1.08	1.05	1.04	1.02	1.01	1.01	1.01	1.01	1.01
$K = 9$	1.10	1.08	1.07	1.05	1.03	1.02	1.01	1.01	1.01	1.01	1.01

Table 5: Ratio $r_K(\tilde{s}, \tilde{s}_0)$ for the estimation of $s = f_1$ with $\mathcal{F}^N$.

Taking the expectation on both sides, it leads to

$$\begin{aligned}
(1-\alpha)\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] &\leqslant \|s-s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n \\
&\quad +(2-\alpha)\mathbb{E}\left[\sup_{m'\in\mathcal{M}}\left(\sigma^2\|\pi_{m'}P_n\varepsilon\|_n^2 - p_1(m')\right)_+\right] \\
&\quad +\alpha^{-1}\mathbb{E}\left[\sup_{m'\in\mathcal{M}}\left(\sigma^2\langle u_{m'}, P_n\varepsilon\rangle_n^2 - p_2(m')\right)_+\right] \\
&\leqslant \|s-s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n \\
&\quad +(2-\alpha)\sum_{m'\in\mathcal{M}}\mathbb{E}\left[\left(\sigma^2\|\pi_{m'}P_n\varepsilon\|_n^2 - p_1(m')\right)_+\right] \\
&\quad +\alpha^{-1}\sum_{m'\in\mathcal{M}}\mathbb{E}\left[\left(\sigma^2\langle u_{m'}, P_n\varepsilon\rangle_n^2 - p_2(m')\right)_+\right] \\
&\leqslant \|s-s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n \\
&\quad +(2-\alpha)\sum_{m'\in\mathcal{M}}\mathbb{E}_{1,m'} + \alpha^{-1}\sum_{m'\in\mathcal{M}}\mathbb{E}_{2,m'} \ .
\end{aligned}$$

C	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$K=1$	1.08	1.09	1.07	1.07	1.09	1.09	1.08	1.07	1.06	1.09	1.07
$K=2$	1.09	1.05	1.08	1.09	1.09	1.08	1.08	1.08	1.06	1.06	1.05
$K=3$	1.12	1.12	1.11	1.07	1.09	1.10	1.09	1.08	1.07	1.06	1.07
$K=4$	1.09	1.11	1.08	1.10	1.10	1.09	1.07	1.06	1.07	1.06	1.07
$K=5$	1.10	1.08	1.09	1.09	1.09	1.06	1.06	1.06	1.07	1.07	1.05
$K=6$	1.08	1.04	1.06	1.07	1.08	1.07	1.07	1.09	1.06	1.06	1.06
$K=7$	1.06	1.05	1.07	1.08	1.10	1.09	1.07	1.09	1.08	1.07	1.06
$K=8$	1.08	1.13	1.08	1.09	1.09	1.08	1.06	1.07	1.07	1.06	1.06
$K=9$	1.13	1.05	1.09	1.09	1.07	1.07	1.07	1.06	1.07	1.07	1.06

Table 6: Ratio $r_K(\tilde{s},\tilde{s}_0)$ for the estimation of $s=f_5$ with $\mathcal{F}^N$.

Because the choice of m is arbitrary among $\mathcal{M}$, we can infer that

$$(1-\alpha)\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2+\mathrm{pen}(m)-\mathrm{Tr}({}^tP_n\pi_mP_n)\sigma^2/n\right\} +(2-\alpha)\sum_{m\in\mathcal{M}}\mathbb{E}_{1,m}+\alpha^{-1}\sum_{m\in\mathcal{M}}\mathbb{E}_{2,m}\ . \tag{41}$$

We now have to upperbound $\mathbb{E}_{1,m}$ and $\mathbb{E}_{2,m}$ in (41). Let start by the first one. If $S_m=\{0\}$, then $\pi_mP_n=0$ and $p_1(m)\geqslant 0$ suffices to ensure that $\mathbb{E}_{1,m}=0$. So, we can consider that the dimension of S_m is positive and $\pi_mP_n\neq 0$. The Lemma 8.2 applied with $A=\pi_mP_n$ gives, for any $x>0$,

$$\mathbb{P}\left(n\|\pi_mP_n\varepsilon\|_n^2\geqslant \mathrm{Tr}({}^tP_n\pi_mP_n)+2\sqrt{\rho^2(P_n)\mathrm{Tr}({}^tP_n\pi_mP_n)x}+2\rho^2(P_n)x\right)\leqslant e^{-x} \tag{42}$$

because $\rho(\pi_mP_n)\leqslant\rho(\pi_m)\rho(P_n)\leqslant\rho(P_n)$. Let $\beta=\theta^2/(1+2\theta)>0$, (35) and (42) lead to

$$\mathbb{P}\left(n\|\pi_mP_n\varepsilon\|_n^2\geqslant(1+\beta)\mathrm{Tr}({}^tP_n\pi_mP_n)+(2+\beta^{-1})\rho^2(P_n)x\right)\leqslant e^{-x}\ . \tag{43}$$

Let $\delta=\theta^2/((1+\theta)(1+2\theta+2\theta^2))>0$, we set

$$np_1(m)=((1+\beta)+(2+\beta^{-1})\delta L_m)\mathrm{Tr}({}^tP_n\pi_mP_n)\sigma^2$$

and (43) implies

$$\begin{aligned}
\mathbb{E}_{m,1} &= \int_0^\infty \mathbb{P}\left(\left(\sigma^2\|\pi_mP_n\varepsilon\|_n^2-p_1(m)\right)_+\geqslant\xi\right)d\xi\\
&= \int_0^\infty \mathbb{P}\left(n\|\pi_mP_n\varepsilon\|_n^2-np_1(m)/\sigma^2\geqslant n\xi/\sigma^2\right)d\xi\\
&\leqslant \int_0^\infty \exp\left(-\frac{\delta L_m\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}-\frac{n\xi}{(2+\beta^{-1})\rho^2(P_n)\sigma^2}\right)d\xi\\
&\leqslant \frac{(2+\beta^{-1})\rho^2(P_n)\sigma^2}{n}\exp\left(-\frac{\delta L_m\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}\right)\ .
\end{aligned} \tag{44}$$

We now focus on $\mathbb{E}_{m,2}$. The random variable $\langle u_m,P_n\varepsilon\rangle_n=\langle {}^tP_nu_m,\varepsilon\rangle_n$ is a centered Gaussian variable with variance $\|{}^tP_nu_m\|_n^2/n$. For any $x>0$, the standard Gaussian deviation

inequality gives

$$\mathbb{P}\left(|\langle u_m, P_n\varepsilon\rangle_n| \geqslant x\right) \leqslant \exp\left(-\frac{nx^2}{2\|\,{}^tP_n u_m\|_n^2}\right) \leqslant \exp\left(-\frac{nx^2}{2\rho^2(P_n)}\right)$$

that is equivalent to

$$\mathbb{P}\left(n\langle u_m, P_n\varepsilon\rangle_n^2 \geqslant 2\rho^2(P_n)x\right) \leqslant e^{-x}\ . \tag{45}$$

We set

$$np_2(m) = 2\delta L_m \mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2$$

and (45) leads to

$$\begin{aligned}
\mathbb{E}_{m,2} &= \int_0^\infty \mathbb{P}\left(\left(\sigma^2\langle u_m, P_n\varepsilon\rangle_n^2 - p_2(m)\right)_+ \geqslant \xi\right) d\xi \\
&= \int_0^\infty \mathbb{P}\left(\langle u_m, P_n\varepsilon\rangle_n^2 - np_2(m)/\sigma^2 \geqslant n\xi/\sigma^2\right) d\xi \\
&\leqslant \int_0^\infty \exp\left(-\frac{\delta L_m \mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho^2(P_n)} - \frac{n\xi}{2\rho^2(P_n)\sigma^2}\right) d\xi \\
&\leqslant \frac{2\rho^2(P_n)\sigma^2}{n}\exp\left(-\frac{\delta L_m \mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho^2(P_n)}\right)\ .
\end{aligned} \tag{46}$$

We inject (44) and (46) in (41) and we replace α, β and δ to obtain

$$\frac{\theta}{\theta+1}\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + \mathrm{pen}(m) - \mathrm{Tr}({}^tP_n\pi_m P_n)\sigma^2/n\right\} + \frac{\rho^2(P_n)\sigma^2}{n}R_\theta$$

where we have set

$$R_\theta = c_\theta \sum_{m\in\mathcal{M}} \exp\left(-\frac{L_m \mathrm{Tr}({}^tP_n\pi_m P_n)}{c_\theta\rho^2(P_n)}\right)$$

and

$$c_\theta = \frac{2\theta^4+8\theta^3+8\theta^2+4\theta+1}{\theta^2(1+\theta)}\ .$$

Finally, (40) gives a penalty as (15) and the announced result follows.

7.1.2 Proof of Theorem 2.2

In order to prove Theorem 2.2, we show the following stronger result. Under the assumptions of the theorem, there exists a positive constant C that only depends on p and θ, such that, for any $z>0$,

$$\mathbb{P}\left(\frac{\theta}{\theta+2}\mathcal{H}_+ \geqslant \frac{\rho^2(P_n)\sigma^2}{n}z\right) \leqslant C\tau_p\left[N_0\left(1\wedge z^{-p/2}\right) + R_{P_n,p}(\mathcal{F},z)\right] \tag{47}$$

where the quantity $\mathcal{H}$ is defined by

$$\mathcal{H} = \|s-\tilde{s}\|_n^2 - \frac{\theta+4}{\theta}\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + \frac{2(\theta+2)}{\theta+4}\mathrm{pen}(m)\right\}$$

and we have set $R_{P_n,p}(\mathcal{F},z)$ equal to

$$\sum_{m\in\mathcal{M}:S_m\neq\{0\}}\left(1+\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho({}^tP_n\pi_mP_n)}\right)\left(\frac{L_m\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}+z\right)^{-p/2}.$$

Thus, for any $q>0$ such that $2(q+1)<p$, we integrate (47) via Lemma 8.1 to get

$$\begin{aligned}\mathbb{E}\left[\mathcal{H}_+^q\right] &= \int_0^\infty qt^{q-1}\mathbb{P}\left(\mathcal{H}_+\geqslant t\right)dt\\ &= \left(\frac{(\theta+2)\rho^2(P_n)\sigma^2}{\theta n}\right)^q\int_0^\infty qz^{q-1}\mathbb{P}\left(\frac{\theta}{\theta+2}\mathcal{H}_+\geqslant\frac{\rho^2(P_n)\sigma^2}{n}z\right)dz\\ &\leqslant C'(p,q,\theta)\tau_p\left(\frac{\rho^2(P_n)\sigma^2}{n}\right)^q R^{p,q}_{P_n,\theta}(\mathcal{F}) \qquad (48)\end{aligned}$$

where we have set

$$R^{p,q}_{P_n,\theta}(\mathcal{F})=N_0+\sum_{m\in\mathcal{M}:S_m\neq\{0\}}\left(1+\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho({}^tP_n\pi_mP_n)}\right)\left(\frac{L_m\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}\right)^{q-p/2}.$$

Since

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]^{1/q}\leqslant\mathbb{E}\left[\left(\frac{\theta+8}{\theta}\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2+\frac{2(\theta+4)}{\theta+8}\mathrm{pen}(m)\right\}+\mathcal{H}_+\right)^q\right]^{1/q},$$

it follows from Minkowski's Inequality when $q\geqslant 1$ or convexity arguments when $0<q<1$ that

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\right]^{1/q}\leqslant 2^{(q^{-1}-1)_+}\left(C''(\theta)\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2+\mathrm{pen}(m)\right\}+\mathbb{E}\left[\mathcal{H}_+^q\right]^{1/q}\right). \qquad (49)$$

Inequality (18) directly follows from (48) and (49).

We now turn to the proof of (47). Inequality (39) does not depend on the distribution of ε and we start from here. Let $\alpha=\alpha(\theta)\in(0,1)$, for any $m\in\mathcal{M}$ we have

$$\begin{aligned}(1-\alpha)\|s-\tilde{s}\|_n^2 &\leqslant \|s-s_m\|_n^2+\mathrm{pen}(m)-\mathrm{pen}(\hat{m})+(2-\alpha)\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2\\ &\quad+\alpha^{-1}\sigma^2\langle u_{\hat{m}},P_n\varepsilon\rangle_n^2+2\sigma\langle s-s_m,P_n\varepsilon\rangle_n\end{aligned}$$

where u_m is defined by (37). Use again (35) with α to obtain

$$\begin{aligned}(1-\alpha)\|s-\tilde{s}\|_n^2 &\leqslant \|s-s_m\|_n^2+\mathrm{pen}(m)-\mathrm{pen}(\hat{m})+(2-\alpha)\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2\\ &\quad+\alpha^{-1}\sigma^2\langle u_{\hat{m}},P_n\varepsilon\rangle_n^2+2\sigma\|s-s_m\|_n|\langle u_m,P_n\varepsilon\rangle_n|\\ &\leqslant (1+\alpha)\|s-s_m\|_n^2+\mathrm{pen}(m)-\mathrm{pen}(\hat{m}) \qquad (50)\\ &\quad+(2-\alpha)\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2\\ &\quad+\alpha^{-1}\sigma^2\langle u_{\hat{m}},P_n\varepsilon\rangle_n^2+\alpha^{-1}\sigma^2\langle u_m,P_n\varepsilon\rangle_n^2.\end{aligned}$$

Let us now introduce two functions $\bar{p}_1,\bar{p}_2:\mathcal{M}\to\mathbb{R}_+$ that will be specified later and that satisfy,

$$\forall m\in\mathcal{M},\ \mathrm{pen}(m)\geqslant(2-\alpha)\bar{p}_1(m)+\alpha^{-1}\bar{p}_2(m). \qquad (51)$$

Thus, Inequality (50) implies

$$\begin{aligned}(1-\alpha)\|s-\tilde{s}\|_n^2 &\leqslant (1+\alpha)\|s-s_m\|_n^2+\mathrm{pen}(m)+\alpha^{-1}\bar{p}_2(m)\\ &\quad+(2-\alpha)\left(\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2-\bar{p}_1(\hat{m})\right)\\ &\quad+\alpha^{-1}\left(\sigma^2\langle u_{\hat{m}},P_n\varepsilon\rangle_n^2-\bar{p}_2(\hat{m})\right)\\ &\quad+\alpha^{-1}\left(\sigma^2\langle u_m,P_n\varepsilon\rangle_n^2-\bar{p}_2(m)\right)\\ &\leqslant (1+\alpha)\left(\|s-s_m\|_n^2+2\mathrm{pen}(m)/(1+\alpha)\right)\\ &\quad+(2-\alpha)\sup_{m'\in\mathcal{M}}\left(\sigma^2\|\pi_{m'}P_n\varepsilon\|_n^2-\bar{p}_1(m')\right)_+\\ &\quad+2\alpha^{-1}\sup_{m'\in\mathcal{M}}\left(\sigma^2\langle u_{m'},P_n\varepsilon\rangle_n^2-\bar{p}_2(m')\right)_+ .\end{aligned}$$

Because the choice of m is arbitrary among $\mathcal{M}$, we can infer that, for any $\xi>0$,

$$\begin{aligned}\mathbb{P}\left((1-\alpha)\mathcal{H}_+\geqslant\xi\right) &\leqslant \mathbb{P}\left((2-\alpha)\sup_{m\in\mathcal{M}}\left(\sigma^2\|\pi_mP_n\varepsilon\|_n^2-\bar{p}_1(m)\right)_+\geqslant\frac{\xi}{2}\right)\\ &\quad+\mathbb{P}\left(2\alpha^{-1}\sup_{m\in\mathcal{M}}\left(\sigma^2\langle u_m,P_n\varepsilon\rangle_n^2-\bar{p}_2(m)\right)_+\geqslant\frac{\xi}{2}\right)\\ &\leqslant \sum_{m\in\mathcal{M}}\mathbb{P}\left(\sigma^2\|\pi_mP_n\varepsilon\|_n^2\geqslant\bar{p}_1(m)+\frac{\xi}{2(2-\alpha)}\right)\\ &\quad+\sum_{m\in\mathcal{M}}\mathbb{P}\left(\sigma^2\langle u_m,P_n\varepsilon\rangle_n^2\geqslant\bar{p}_2(m)+\frac{\alpha\xi}{4}\right)\\ &\leqslant \sum_{m\in\mathcal{M}}\mathbb{P}_{1,m}(\xi)+\sum_{m\in\mathcal{M}}\mathbb{P}_{2,m}(\xi) . \qquad (52)\end{aligned}$$

We first bound $\mathbb{P}_{1,m}(\xi)$. For $m\in\mathcal{M}$ such that $S_m=\{0\}$ (*i.e.* $\pi_m=0$), $\bar{p}_1(m)\geqslant 0$ leads obviously to $\mathbb{P}_{1,m}(\xi)=0$. Thus, it is sufficient to bound $\mathbb{P}_{1,m}(\xi)$ for m such that π_m is not null. This ensures that the symmetric nonnegative matrix $\tilde{A}={}^tP_n\pi_mP_n$ lies in $\mathbb{M}_n\setminus\{0\}$. Thus, under hypothesis (6), Corollary 5.1 of [4] gives us, for any $x_m>0$,

$$\mathbb{P}\left(n\|\pi_mP_n\varepsilon\|_n^2\geqslant\mathrm{Tr}(\tilde{A})+2\sqrt{\rho(\tilde{A})\mathrm{Tr}(\tilde{A})x_m}+\rho(\tilde{A})x_m\right)\leqslant\frac{C_1(p)\tau_p\mathrm{Tr}(\tilde{A})}{\rho(\tilde{A})x_m^{p/2}}$$

where $C_1(p)$ is a constant that only depends on p. The properties of the norm ρ imply

$$\rho(\tilde{A})=\rho({}^t(\pi_mP_n)(\pi_mP_n))=\rho(\pi_mP_n)^2\leqslant\rho^2(P_n) . \qquad (53)$$

By the inequalities (53) and (35) with $\theta/2>0$, we obtain

$$\mathbb{P}\left(n\|\pi_mP_n\varepsilon\|_n^2\geqslant\left(1+\frac{\theta}{2}\right)\mathrm{Tr}(\tilde{A})+\left(1+\frac{2}{\theta}\right)\rho^2(P_n)x_m\right)\leqslant\frac{C_1(p)\tau_p\mathrm{Tr}(\tilde{A})}{\rho(\tilde{A})x_m^{p/2}} . \qquad (54)$$

We take $\alpha=2/(\theta+2)\in(0,1)$ but for legibility we keep using the notation α. Moreover, we choose

$$n\bar{p}_1(m)=\left(1+\frac{\theta}{2}+\frac{L_m}{2(\theta+1)}\right)\mathrm{Tr}({}^tP_n\pi_mP_n)\sigma^2$$

and

$$x_m = \frac{\theta}{2(\theta+1)(\theta+2)} \times \frac{L_m \mathrm{Tr}({}^tP_n\pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)} \ .$$

Thus, Inequality (54) leads to

$$\begin{aligned}
\mathbb{P}_{1,m}(\xi) &= \mathbb{P}\left(\sigma^2\|\pi_m P_n\varepsilon\|_n^2 \geqslant \bar{p}_1(m) + \frac{\xi}{2(2-\alpha)}\right) \\
&= \mathbb{P}\left(\sigma^2\|\pi_m P_n\varepsilon\|_n^2 \geqslant \bar{p}_1(m) + \frac{(\theta+2)\xi}{4(\theta+1)}\right) \\
&\leqslant \mathbb{P}\left(n\|\pi_m P_n\varepsilon\|_n^2 \geqslant \left(1+\frac{\theta}{2}\right)\mathrm{Tr}({}^tP_n\pi_m P_n) + \left(1+\frac{2}{\theta}\right)\rho^2(P_n)x_m\right) \\
&\leqslant C_2(p,\theta)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)\tau_p}{\rho({}^tP_n\pi_m P_n)}\left(\frac{L_m\mathrm{Tr}({}^tP_n\pi_m P_n)+n\xi/\sigma^2}{\rho^2(P_n)}\right)^{-p/2} \ . \qquad (55)
\end{aligned}$$

We now focus on $\mathbb{P}_{2,m}(\xi)$. Let y_m be some positive real number, the Markov Inequality leads to

$$\mathbb{P}\left(|\langle u_m, P_n\varepsilon\rangle_n| \geqslant y_m\right) \leqslant y_m^{-p}\mathbb{E}\left[|\langle u_m, P_n\varepsilon\rangle_n|^p\right] = y_m^{-p}\mathbb{E}\left[\left|\langle {}^tP_n u_m, \varepsilon\rangle_n\right|^p\right] \ . \qquad (56)$$

Since $p > 2$, the quantity τ_p is lower bounded by 1,

$$\tau_p = \mathbb{E}\left[|\varepsilon_1|^p\right] \geqslant \mathbb{E}\left[\varepsilon_1^2\right]^{p/2} = 1 \ . \qquad (57)$$

Moreover, we can apply the Rosenthal inequality (see Chapter 2 of [31]) to obtain

$$\mathbb{E}\left[\left|\langle {}^tP_n u_m, \varepsilon\rangle_n\right|^p\right] \leqslant C_3(p)n^{-p}\left(\tau_p\sum_{i=1}^n\left|({}^tP_n u_m)_i\right|^p + n^{p/2}\|{}^tP_n u_m\|_n^p\right) \qquad (58)$$

where $C_3(p)$ is a constant that only depends on p. Since $p > 2$, we have

$$\sum_{i=1}^n\left|({}^tP_n u_m)_i\right|^p \leqslant \left(\sum_{i=1}^n({}^tP_n u_m)_i^2\right)^{p/2} = n^{p/2}\|{}^tP_n u_m\|_n^p \leqslant n^{p/2}\rho^p(P_n) \ .$$

Thus, the Inequality (58) becomes

$$\mathbb{E}\left[\left|\langle {}^tP_n u_m, \varepsilon\rangle_n\right|^p\right] \leqslant 2C_3(p)\rho^p(P_n)\tau_p n^{-p/2}$$

and, putting this inequality in (56), we obtain

$$\mathbb{P}\left(|\langle u_m, P_n\varepsilon\rangle_n| \geqslant y_m\right) \leqslant 2C_3(p)\rho^p(P_n)\tau_p n^{-p/2}y_m^{-p} \ . \qquad (59)$$

We take

$$n\bar{p}_2(m) = \frac{1}{2(\theta+1)}\sigma^2 L_m\mathrm{Tr}({}^tP_n\pi_m P_n)$$

and

$$y_m^2 = \frac{1}{2(\theta+2)n}\left(L_m\mathrm{Tr}({}^tP_n\pi_m P_n) + \frac{n\xi}{\sigma^2}\right) \ .$$

Finally, (59) gives

$$\begin{aligned}
\mathbb{P}_{2,m}(\xi) &= \mathbb{P}\left(\sigma^2\langle u_m, P_n\varepsilon\rangle_n^2 \geqslant \bar{p}_2(m) + \frac{\alpha\xi}{4}\right) \\
&= \mathbb{P}\left(\sigma^2\langle u_m, P_n\varepsilon\rangle_n^2 \geqslant \bar{p}_2(m) + \frac{\xi}{2(\theta+2)}\right) \\
&\leqslant \mathbb{P}\left(\langle u_m, P_n\varepsilon\rangle_n^2 \geqslant y_m^2\right) \\
&\leqslant C_4(p,\theta)\tau_p\left(\frac{L_m \mathrm{Tr}({}^tP_n\pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)}\right)^{-p/2} . \qquad (60)
\end{aligned}$$

Taking

$$R'(\xi) = \sum_{m\in\mathcal{M}:S_m\neq\{0\}} \left(1 + \frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho({}^tP_n\pi_m P_n)}\right)\left(\frac{L_m \mathrm{Tr}({}^tP_n\pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)}\right)^{-p/2}$$

and putting together Inequalities (52), (55) and (60) lead us to

$$\begin{aligned}
\mathbb{P}\left((1-\alpha)\mathcal{H}_+ \geqslant \xi\right) &\leqslant \sum_{m\in\mathcal{M}} \mathbb{P}_{1,m}(\xi) + \sum_{m\in\mathcal{M}} \mathbb{P}_{2,m}(\xi) \\
&\leqslant \sum_{m\in\mathcal{M}:S_m=\{0\}} \mathbb{P}_{2,m}(\xi) + \sum_{m\in\mathcal{M}:S_m\neq\{0\}} \mathbb{P}_{1,m}(\xi) + \sum_{m\in\mathcal{M}:S_m\neq\{0\}} \mathbb{P}_{2,m}(\xi) \\
&\leqslant \sum_{m\in\mathcal{M}:S_m=\{0\}} 1 \wedge \left\{C_4(p,\theta)\tau_p\left(\frac{n\xi}{\sigma^2\rho^2(P_n)}\right)^{-p/2}\right\} + C_5(p,\theta)\tau_p R'(\xi) \\
&\leqslant N_0(1 \vee C_4(p\theta))\tau_p\left(1 \wedge \left(\frac{n\xi}{\rho^2(P_n)\sigma^2}\right)^{-p/2}\right) + C_5(p,\theta)\tau_p R'(\xi) .
\end{aligned}$$

For $z > 0$, take $\xi = \rho^2(P_n)\sigma^2 z/n$ to obtain (47). We conclude the proof by computing the lowerbound (51) on the penalty function,

$$\begin{aligned}
(2-\alpha)\bar{p}_1(m) + \alpha^{-1}\bar{p}_2(m) &= \frac{2(\theta+1)}{\theta+2}\bar{p}_1(m) + \frac{\theta+2}{2}\bar{p}_2(m) \\
&= \left(1 + \theta + \frac{\theta^2+8\theta+8}{4(\theta+1)(\theta+2)}L_m\right)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 .
\end{aligned}$$

Since $(\theta^2+8\theta+8)/(4(\theta+1)(\theta+2)) \leqslant 1$, the penalty given by (17) satisfies the condition (51).

7.2 Proofs of Theorems 3.1 and 3.2

7.2.1 Proof of Theorem 3.1

Given $\theta > 0$, we can find two positive numbers $\delta = \delta(\theta) < 1/2$ and $\eta = \eta(\theta)$ such that $(1+\theta)(1-2\delta) \geqslant (1+2\eta)$. Thus we define

$$\Omega_n = \left\{\hat{\sigma}^2 > (1-2\delta)\sigma^2\right\} .$$

On Ω_n, we know that

$$\forall m \in \mathcal{M},\ \mathrm{pen}(m) \geqslant (1+2\eta)\frac{Tr({}^tP_n\pi_m P_n)}{n}\sigma^2\ .$$

Taking care of the random nature of the penalty, we argue as in the proof of Theorem 2.1 with $L_m = \eta$ to get

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_n}\right] \leqslant \frac{\eta+1}{\eta}\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + \mathbb{E}[\mathrm{pen}(m)] - \frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2\right\} + \frac{\rho^2(P_n)\sigma^2}{n}R''_{P_n,\eta}(\mathcal{F}) \tag{61}$$

where $R''_{P_n,\eta}(\mathcal{F})$ is defined by

$$R''_{P_n,\eta}(\mathcal{F}) = C_\eta \sum_{m\in\mathcal{M}} \exp\left(-\frac{C'_\eta \mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho^2(P_n)}\right)\ .$$

We use Lemma 8.3 and (21) to get an upperbound for $\mathbb{E}[\mathrm{pen}(m)]$,

$$\begin{aligned}
\mathbb{E}[\mathrm{pen}(m)] &\leqslant (1+\theta)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 + (1+\theta)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)\|s-\pi s\|_n^2}{\mathrm{Tr}\left({}^tP_n(I_n-\pi)P_n\right)} \\
&\leqslant (1+\theta)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 + (1+\theta)\frac{\mathrm{Tr}({}^tP_n P_n)\|s-\pi s\|_n^2}{\mathrm{Tr}\left({}^tP_n(I_n-\pi)P_n\right)} \\
&\leqslant (1+\theta)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2 + 2(1+\theta)\|s-\pi s\|_n^2\ .
\end{aligned}$$

The Proposition 1.1 and (61) give

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_n}\right] \leqslant C(\theta)\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right] + 2(\theta+1)\|s-\pi s\|_n^2 + \frac{\rho^2(P_n)\sigma^2}{n}R''_{P_n,\eta}(\mathcal{F}) \tag{62}$$

where $C(\theta) > 1$.

We now bound $\mathbb{E}[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}]$. Note that

$$\|s-\tilde{s}\|_n^2 = \|s-s_{\hat{m}}\|_n^2 + \sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 \leqslant \|s\|_n^2 + \sigma^2\|P_n\varepsilon\|_n^2$$

and thus, by the Cauchy–Schwarz Inequality,

$$\mathbb{E}[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}] \leqslant \|s\|_n^2\mathbb{P}\left(\Omega_n^c\right) + \sigma^2\mathbb{E}[\|P_n\varepsilon\|_n^2 \mathbb{1}_{\Omega_n^c}] \leqslant \left(\|s\|_n^2 + \sigma^2\mathbb{E}[\|P_n\varepsilon\|_n^4]^{1/2}\right)\mathbb{P}\left(\Omega_n^c\right)^{1/2}\ .$$

Moreover, the eigenvalues of the matrix $P_n\,{}^tP_n$ are nonnegative and so

$$\begin{aligned}
\mathbb{E}[\|P_n\varepsilon\|_n^4]^{1/2} &= \left(\mathrm{Var}(\|P_n\varepsilon\|_n^2) + E[\|P_n\varepsilon\|_n^2]^2\right)^{1/2} \\
&\leqslant \frac{1}{n}\sqrt{\mathrm{Tr}({}^tP_nP_n)\left(\mathrm{Tr}({}^tP_nP_n) + 2\rho^2(P_n)\right)} \\
&\leqslant \frac{\mathrm{Tr}({}^tP_nP_n) + \left(\mathrm{Tr}({}^tP_nP_n) + 2\rho^2(P_n)\right)}{2n} \\
&\leqslant \frac{\mathrm{Tr}({}^tP_nP_n) + \rho^2(P_n)}{n}\ .
\end{aligned}$$

Finally, the Lemma 8.4 gives

$$\begin{aligned}\mathbb{E}[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}] &\leqslant C'(\theta)\left(\|s\|_n^2+\frac{\mathrm{Tr}({}^tP_nP_n)+\rho^2(P_n)}{n}\sigma^2\right)\exp\left(-\frac{\theta^2\mathrm{Tr}({}^tP_nP_n)}{32\rho^2(P_n)}\right)\\ &\leqslant C'(\theta)\left(\|s\|_n^2+\frac{\rho^2(P_n)(n+1)}{n}\sigma^2\right)\exp\left(-\frac{\theta^2\mathrm{Tr}({}^tP_nP_n)}{32\rho^2(P_n)}\right)\\ &\leqslant C'(\theta)\left(\|s\|_n^2+2\rho^2(P_n)\sigma^2\right)\exp\left(-\frac{\theta^2\mathrm{Tr}({}^tP_nP_n)}{32\rho^2(P_n)}\right) \qquad (63)\end{aligned}$$

where $C'(\theta)>1$. The inequality (24) follows by collecting (62) and (63).

7.2.2 Proof of Theorem 3.2

Given $\theta>0$, we can find two positive numbers $\delta=\delta(\theta)<1/3$ and $\eta=\eta(\theta)$ such that $(1+\theta)(1-3\delta)\geqslant(1+2\eta)$. Thus we define

$$\Omega_n'=\left\{\hat{\sigma}^2>(1-3\delta)\sigma^2\right\}\ .$$

On Ω_n', we know that

$$\forall m\in\mathcal{M},\ \mathrm{pen}(m)\geqslant(1+2\eta)\frac{Tr({}^tP_n\pi_mP_n)}{n}\sigma^2\ .$$

Let $\bar{m}$ be any element of $\mathcal{M}$ that minimize $\|s-s_{m'}\|_n^2+\sigma^2\mathrm{Tr}({}^tP_n\pi_{m'}P_n)/n$ among $m'\in\mathcal{M}$. Taking care of the random nature of the penalty, we argue as in the proof of Theorem 2.2 with $L_m=\eta$ to get

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega_n'}\right]^{1/q}\leqslant C(q,\theta)\mathbb{E}\left[\left(\|s-s_{\bar{m}}\|_n^2+\frac{\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)}{n}\hat{\sigma}^2\right)^q\right]^{1/q}+\frac{\rho^2(P_n)\sigma^2}{n}R_n(p,q,\theta)^{1/q}$$

where $R_n(p,q,\theta)$ is equal to

$$C'(p,q,\theta)\tau_p\left[N_0+\sum_{m\in\mathcal{M}:S\neq\{0\}}\left(1+\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho({}^tP_n\pi_mP_n)}\right)\left(\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}\right)^{q-p/2}\right]\ .$$

Since $q\leqslant 1$, by a convexity argument and Jensen's inequality we deduce

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega_n'}\right]^{1/q}\leqslant C(q,\theta)\left(\|s-s_{\bar{m}}\|_n^2+\frac{\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)}{n}\mathbb{E}[\hat{\sigma}^2]\right)+\frac{\rho^2(P_n)\sigma^2}{n}R_n(p,q,\theta)^{1/q}\ . \qquad (64)$$

Lemma 8.3 and (21) give

$$\begin{aligned}\frac{\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)}{n}\mathbb{E}[\hat{\sigma}^2] &= \frac{\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)}{n}\sigma^2+\frac{n\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)\|s-\pi s\|_n^2}{n\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)}\\ &\leqslant \frac{\mathrm{Tr}({}^tP_n\pi_{\bar{m}}P_n)}{n}\sigma^2+2\|s-\pi s\|_n^2\ .\end{aligned}$$

Thus, by the definition of $\bar{m}$ and Proposition 1.1, (64) becomes

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega_n'}\right]^{1/q}\leqslant C(q,\theta)\left(\inf_{m\in\mathcal{M}}\mathbb{E}[\|s-\hat{s}_m\|_n^2]+2\|s-\pi s\|_n^2\right)+\frac{\rho^2(P_n)\sigma^2}{n}R_n(p,q,\theta)^{1/q}\ . \qquad (65)$$

We now bound $\mathbb{E}[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega'^c_n}]$. Note that

$$\|s-\tilde{s}\|_n^2 = \|s-s_{\hat{m}}\|_n^2 + \sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 \leqslant \|s\|_n^2 + \sigma^2\|P_n\varepsilon\|_n^2 \ .$$

Since $q \leqslant 1$, we have

$$\mathbb{E}[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega'^c_n}] \leqslant \|s\|_n^{2q}\mathbb{P}(\Omega'^c_n) + \sigma^{2q}\mathbb{E}[\|P_n\varepsilon\|_n^{2q}\mathbb{1}_{\Omega'^c_n}] \ .$$

Hölder's Inequality with exponent $p/2q > 1$ gives

$$\mathbb{E}[\|P_n\varepsilon\|_n^{2q}\mathbb{1}_{\Omega'^c_n}] \leqslant \mathbb{E}[\|P_n\varepsilon\|_n^p]^{2q/p}\mathbb{P}(\Omega'^c_n)^{1-2q/p}$$

and, since

$$\mathbb{E}[\|P_n\varepsilon\|_n^p]^{2q/p} \leqslant \rho^{2q}(P_n)\mathbb{E}[\|\varepsilon\|_n^p]^{2q/p} \leqslant \rho^{2q}(P_n)\tau_p^{2q/p} \ ,$$

we obtain by using Lemma 8.5 that

$$\begin{aligned}\mathbb{E}[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega'^c_n}] &\leqslant (\|s\|_n^{2q} + \sigma^{2q}\rho^{2q}(P_n)\tau_p^{2q/p})\mathbb{P}(\Omega'^c_n)^{1-2q/p} \\ &\leqslant C(p,q,\theta)\kappa'_n(p,q,\theta)(\|s\|_n^{2q} + \sigma^{2q}\rho^{2q}(P_n)\tau_p^{2q/p})\left(\tau_p\rho^{\alpha_p}(P_n)\mathrm{Tr}({}^tP_nP_n)^{-\beta_p}\right)^{1-2q/p}\end{aligned}$$

where

$$\alpha_p = (p/2-1)\vee 1 \text{ and } \beta_p = (p/2-1)\wedge 1 \ .$$

Thus, we get

$$\mathbb{E}[\|s-\tilde{s}\|_n^{2q}\mathbb{1}_{\Omega'^c_n}]^{1/q} \leqslant C'(p,q,\theta)\kappa_n(p,q,\theta)\tau_p^{1/q}(\|s\|_n^2 + \tau_p\rho^2(P_n)\sigma^2)\left(\frac{\rho^{2\alpha_p}(P_n)}{\mathrm{Tr}({}^tP_nP_n)^{\beta_p}}\right)^{1/q-2/p} . \tag{66}$$

The announced result follows from (65) and (66).

7.3 Proofs of Corollaries and Propositions

7.3.1 Proof of Corollary 2.1

Let us begin by applying Theorem 2.1 with constant weights $L_m = L$,

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant (1+\theta^{-1})\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (\theta+L)\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{n}\sigma^2\right\} + \frac{\rho^2(P_n)\sigma^2}{n}R_n(\theta) \ . \tag{67}$$

We now upperbound the remainder term. Assumption $(\mathbf{A}'_3)$ and bounds on N_d and L lead to

$$\begin{aligned}R_n(\theta) &\leqslant \frac{2(1+\theta)^4}{\theta^3}\sum_{m\in\mathcal{M}}\exp\left(-\frac{\theta^2L}{2(1+\theta)^3}\times\frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{\rho^2(P_n)}\right) \\ &\leqslant \frac{2(1+\theta)^4}{\theta^3}\sum_{m\in\mathcal{M}}\exp\left(-\frac{c\theta^2L}{2(1+\theta)^3}\dim(S_m)\right) \\ &\leqslant \frac{2(1+\theta)^4}{\theta^3}\sum_{d\in\mathbb{N}}N_de^{-(A+\omega)d} \\ &\leqslant \frac{2(1+\theta)^4}{\theta^3}\sum_{d\in\mathbb{N}}e^{-\omega d} \ .\end{aligned}$$

The last bound is clearly finite and we denote it by $R = R(\theta, \omega)$. Thus, we derive from (67)

$$\frac{\theta}{\theta+1}\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + \left((\theta+L)\mathrm{Tr}({}^tP_n\pi_m P_n) + R\rho^2(P_n)(\dim(S_m)\vee 1)\right)\frac{\sigma^2}{n}\right\}$$

and hypothesis $(\mathbf{A}_3')$ gives

$$\frac{\theta}{\theta+1}\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (\theta+L+R/c)\left(\mathrm{Tr}({}^tP_n\pi_m P_n)\vee c\rho^2(P_n)\right)\frac{\sigma^2}{n}\right\}$$

that concludes the proof.

7.3.2 Proof of Corollary 2.2

Since $p > 6$, we can take $q = 1$ and apply Theorem 2.2 with constant weights $L_m = L$ to get

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C\inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (1+\theta+L)\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{n}\sigma^2\right\} + \frac{\rho^2(P_n)\sigma^2}{n}R_n(p,1,\theta)\ . \tag{68}$$

To upperbound the remainder term, we use Assumption $(\mathbf{A}_3')$ and bounds on N_d and L to get

$$\begin{aligned}
R_n(p,1,\theta) &\leqslant C'\tau_p\left[1+\sum_{m\in\mathcal{M}:S_m\neq\{0\}}\left(1+\frac{\mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho({}^tP_n\pi_m P_n)}\right)\left(\frac{L\mathrm{Tr}({}^tP_n\pi_m P_n)}{\rho^2(P_n)}\right)^{1-p/2}\right]\\
&\leqslant C'\tau_p\left[1+\sum_{m\in\mathcal{M}:S_m\neq\{0\}}(1+\dim(S_m))(Lc\dim(S_m))^{1-p/2}\right]\\
&\leqslant C'\tau_p\left[1+\frac{(c\omega)'^{1-p/2}}{A}\sum_{d>0}N_d(1+d)d^{1-p/2}\right]\\
&\leqslant C'\tau_p\left[1+(c\omega)'^{1-p/2}\sum_{d>0}(1+d)^{p/2-2-\omega}d^{1-p/2}\right]\ .
\end{aligned}$$

The last bound is clearly finite and we denote it by $R\tau_p = R(\theta, p, \omega, \omega', c)\tau_p$. Thus, as we did in the previous proof, we derive from (68) and $(\mathbf{A}_3')$

$$\frac{1}{C''}\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\|s-s_m\|_n^2 + (1+\theta+L+R\tau_p/c)\left(\mathrm{Tr}({}^tP_n\pi_m P_n)\vee c\rho^2(P_n)\right)\frac{\sigma^2}{n}\right\}\ .$$

Since $\tau_p \geqslant 1$, the announced result follows.

7.3.3 Proof of Proposition 4.1

The design points $(x_i, y_i^1, \ldots, y_i^K)$ are all assumed to be independent realizations of a random variable in $[0,1]^{K+1}$ with distribution $\nu\otimes\nu_1\otimes\cdots\otimes\nu_K$. We denote by I_k the unit $k\times k$ matrix and, for any $a = (a_1, \ldots, a_k)' \in \mathbb{R}^k$, we define the usual norm

$$|a|_2 = \left(\sum_{i=1}^k a_i^2\right)^{1/2}\ .$$

We also consider $\delta_n = \dim(F) \leqslant D_n^{(1)} + \cdots + D_n^{(K)} + 1$ and $N_n = n - D_n - \delta_n$. The quantities δ_n and N_n are random and only depend on the y_i^j's and not on the x_i's.

The space E is generated by the vectors $e^{(i)} = (\phi_i(x_1), \ldots, \phi_i(x_n))'$, for $i = 1, \ldots, D_n$. Let $\{f^{(1)}, \ldots, f^{(\delta_n)}\}$ be an orthonormal basis of F and $\{g^{(1)}, \ldots, g^{(N_n)}\}$ be an orthonormal basis of $G = (E+F)^{\perp}$. In the basis $\mathbf{b}$ of $\mathbb{R}^n$ given by the $e^{(i)}$'s, the $f^{(i)}$'s and the $g^{(i)}$'s, the projection P_n onto E along $F+G$ can be expressed as

$$M = \begin{bmatrix} I_{D_n} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{M}_n(\mathbb{R}) \ .$$

Considering the matrix C that transforms $\mathbf{b}$ into the canonical basis, we can decompose $P_n = CMC^{-1}$. By the properties of the norm ρ, we get

$$\rho^2(P_n) \leqslant \rho^2(C)\rho^2(M)\rho^2(C^{-1}) = \left(\frac{1}{n}\rho({}^tCC)\right)\left(n\rho({}^tC^{-1}C^{-1})\right) \ .$$

For any $\rho > 1$, we deduce from the previous inequality that

$$\mathbb{P}(\rho(P_n) > \rho) \leqslant \mathbb{P}\left(\rho\left(\frac{{}^tCC}{n}\right) > \rho\right) + \mathbb{P}\left(\rho(n\,{}^tC^{-1}C^{-1}) > \rho\right) \ . \tag{69}$$

Note that for any invertible matrix $A \in \mathbb{M}_n(\mathbb{R})$ and $\lambda > 1$, if $\rho(A - I_n) < 1 - \lambda^{-1}$, then $\rho(A^{-1}) < \lambda$. Thus, Inequality (69) leads to

$$\begin{aligned} \mathbb{P}(\rho(P_n) > \rho) &\leqslant \mathbb{P}\left(\rho\left(\frac{{}^tCC}{n}\right) > \rho\right) + \mathbb{P}\left(\rho\left(\frac{{}^tCC}{n} - I_n\right) > 1 - \rho^{-1}\right) \\ &\leqslant 2\mathbb{P}\left(\rho\left(\frac{{}^tCC}{n} - I_n\right) > 1 - \rho^{-1}\right) \ . \end{aligned} \tag{70}$$

Let us denote by Φ the $D_n \times D_n$ Gram matrix associated to the vectors $e^{(1)}, \ldots, e^{(D_n)}$. If we define the $D_n \times \delta_n$ matrix Ω by

$$\forall 1 \leqslant i \leqslant D_n, \ \forall 1 \leqslant j \leqslant \delta_n, \ \Omega_{ij} = \langle e^{(i)}, f^{(j)} \rangle_n \ ,$$

then we can write the following decomposition by blocks,

$$\frac{1}{n}\,{}^tCC = \begin{bmatrix} \Phi & \Omega & 0 \\ {}^t\Omega & I_{\delta_n} & 0 \\ 0 & 0 & I_{N_n} \end{bmatrix} \in \mathbb{M}_n(\mathbb{R}) \ .$$

Consequently, by the definition of $\rho(\cdot)$, we obtain

$$\rho\left(\frac{{}^tCC}{n} - I_n\right) \leqslant \rho(\Phi - I_{D_n}) + \rho(\Omega') \tag{71}$$

where we have set

$$\Omega' = \begin{bmatrix} 0 & \Omega \\ {}^t\Omega & 0 \end{bmatrix} \ .$$

Using (71) in (70) leads to

$$\mathbb{P}(\rho(P_n) > \rho) \leqslant 2\mathbb{P}\left(\rho(\Phi - I_{D_n}) > \frac{1-\rho^{-1}}{2}\right) + 2\mathbb{P}\left(\rho(\Omega') > \frac{1-\rho^{-1}}{2}\right) = 2\mathbb{P}_1 + 2\mathbb{P}_2 \ . \tag{72}$$

First, we upperbound $\mathbb{P}_1$. Let $x > 0$, we consider the event

$$E_x = \left\{ \forall 1 \leqslant i,j \leqslant D_n,\ \left| \langle e^{(i)}, e^{(j)} \rangle_n - \int_0^1 \phi_i(u)\phi_j(u)\nu(du) \right| \leqslant V_{ij}(\phi)\sqrt{2x} + B_{ij}(\phi)x \right\} .$$

Because $\Phi - I_{D_n}$ is symmetric, we know that, on the event E_x,

$$\begin{aligned}
\rho(\Phi - I_{D_n}) &= \sup_{a\in\mathbb{R}^{D_n},\ |a|_2\leqslant 1} |{}^t a(\Phi - I_{D_n})a| \\
&= \sup_{a\in\mathbb{R}^{D_n},\ |a|_2\leqslant 1} \left| \sum_{i=1}^{D_n}\sum_{j=1}^{D_n} a_i a_j \left(\langle e^{(i)}, e^{(j)} \rangle_n - \int_0^1 \phi_i(u)\phi_j(u)\nu(du) \right) \right| 1 \\
&\leqslant \sup_{a\in\mathbb{R}^{D_n},\ |a|_2\leqslant 1} \sum_{i=1}^{D_n}\sum_{j=1}^{D_n} |a_i a_j| \left(|V_{ij}(\phi)|\sqrt{2x} + |B_{ij}(\phi)|x \right) \\
&\leqslant \sqrt{2xL_\phi} + xL_\phi .
\end{aligned}$$

Thus, for any $x > 0$ such that

$$\sqrt{2xL_\phi} + xL_\phi \leqslant \frac{1-\rho^{-1}}{2} \tag{73}$$

we deduce

$$\begin{aligned}
\mathbb{P}_1 &\leqslant \mathbb{P}\left(\exists (i,j)\ :\ \left| \langle e^{(i)}, e^{(j)} \rangle_n - \int_0^1 \phi_i(u)\phi_j(u)\nu(du) \right| > V_{ij}(\phi)\sqrt{2x} + B_{ij}(\phi)x \right) \\
&\leqslant \sum_{i=1}^{D_n}\sum_{j=1}^{D_n} \mathbb{P}\left(\left| \langle e^{(i)}, e^{(j)} \rangle_n - \int_0^1 \phi_i(u)\phi_j(u)\nu(du) \right| > V_{ij}(\phi)\sqrt{2x} + B_{ij}(\phi)x \right) . \qquad (74)
\end{aligned}$$

The choice $x = (1-\rho^{-1})^2/(12L(\phi))$ satisfies (73) and we apply Bernstein Inequality (see Lemma 8 of [8]) to the terms of the sum in (74) to obtain

$$\mathbb{P}_1 \leqslant 2D_n^2 \exp\left(-\frac{n(1-\rho^{-1})^2}{12L_\phi} \right) . \tag{75}$$

It remains to upperbound the probability $\mathbb{P}_2$. Let $x > 0$, we consider the event

$$E'_x = \left\{ \forall 1 \leqslant i \leqslant D_n,\ \forall 1 \leqslant j \leqslant \delta_n,\ \left| \langle e^{(i)}, f^{(j)} \rangle_n \right| \leqslant \sqrt{2x} + b_\phi\sqrt{n}x \right\} .$$

By definition of the norm $\rho(\cdot)$, we know that, on the event E'_x,

$$\begin{aligned}
\rho(\Omega') &= 2 \sup_{\substack{a\in\mathbb{R}^{D_n},\ b\in\mathbb{R}^{\delta_n} \\ |a|_2+|b|_2\leqslant 1}} \left| {}^t a\Omega b \right| \\
&\leqslant 2 \sup_{\substack{a\in\mathbb{R}^{D_n},\ b\in\mathbb{R}^{\delta_n} \\ |a|_2\leqslant 1,\ |b|_2\leqslant 1}} \left| \sum_{i=1}^{D_n}\sum_{j=1}^{\delta_n} a_i b_j \langle e^{(i)}, f^{(j)} \rangle_n \right| \\
&\leqslant 2 \sup_{\substack{a\in\mathbb{R}^{D_n},\ b\in\mathbb{R}^{\delta_n} \\ |a|_2\leqslant 1,\ |b|_2\leqslant 1}} \sum_{i=1}^{D_n}\sum_{j=1}^{\delta_n} |a_i b_j| \left| \langle e^{(i)}, f^{(j)} \rangle_n \right| \\
&\leqslant 2\sqrt{D_n\delta_n} \left(\sqrt{2x} + b_\phi\sqrt{n}x \right) .
\end{aligned}$$

Thus, for any $x > 0$ such that

$$2\sqrt{D_n\delta_n}\left(\sqrt{2x} + b_\phi\sqrt{n}x\right) \leqslant \frac{1-\rho^{-1}}{2}\ , \tag{76}$$

we apply Bernstein Inequality conditionally to the y_i^j's to deduce

$$\begin{aligned}
\mathbb{P}_y\left(\rho(\Omega') > \frac{1-\rho^{-1}}{2}\right) &\leqslant \mathbb{P}_y\left(\exists(i,j)\ :\ \left|\langle e^{(i)}, f^{(j)}\rangle_n\right| > \sqrt{2x} + b_\phi\sqrt{n}x\right) \\
&\leqslant \sum_{i=1}^{D_n}\sum_{j=1}^{\delta_n}\mathbb{P}_y\left(\left|\langle e^{(i)}, f^{(j)}\rangle_n\right| > \sqrt{2x} + b_\phi\sqrt{n}x\right) \\
&\leqslant 2D_n\delta_n e^{-nx} \leqslant 2D_nD_n' e^{-nx}
\end{aligned} \tag{77}$$

where $\mathbb{P}_y$ is the conditional probability given the y_i^j's. Indeed, under $\mathbb{P}_y$ and (30), the variables $\langle e^{(i)}, f^{(j)}\rangle_n$ are centered with unit variance. The choice

$$x = \frac{(1-\rho^{-1})^2}{16\max\left\{4D_n\delta_n, b_\phi\sqrt{nD_n\delta_n}\right\}}$$

satisfies (76) and (77) leads to

$$\begin{aligned}
\mathbb{P}_2 &= \mathbb{E}\left[\mathbb{P}_y\left(\rho(\Omega') > \frac{1-\rho^{-1}}{2}\right)\right] \\
&\leqslant 2D_nD_n'\mathbb{E}\left[\exp\left(-\frac{n(1-\rho^{-1})^2}{16\max\left\{4D_n\delta_n, b_\phi\sqrt{nD_n\delta_n}\right\}}\right)\right] \\
&\leqslant 2D_nD_n'\exp\left(-\frac{n(1-\rho^{-1})^2}{16\max\left\{4D_nD_n', b_\phi\sqrt{nD_nD_n'}\right\}}\right)\ .
\end{aligned} \tag{78}$$

The announced result follows from (72), (75) and (78).

7.3.4 Proof of Proposition 5.1

The collection $\mathcal{F}^{BM}$ is nested and, for any $d \in \mathbb{N}$, the quantity N_d is bounded independently from d. Consequently, Condition (19) is satisfied in the Gaussian case and (20) is fulfilled under moment condition. In both cases, we are free to take $L = \theta = \eta/2$ and $(\mathbf{A}_1)$ is true for $K = \eta$. Assumption $(\mathbf{A}_3')$ is fulfilled with $c = 1/\rho^2$ and, since $\dim(S_m) > 0$ for any $m \in \mathcal{M}$, we can apply Corollary 2.1 or 2.2 according to whether $(\mathbf{H_{Gau}})$ or $(\mathbf{H_{Mom}})$ holds. Moreover, we denote by $\mathbb{E}_\varepsilon$ (*resp.* $\mathbb{E}_d$) the expectation on ε (*resp.* the design points). So $\mathbb{E}_{\varepsilon,d}[\cdot] = \mathbb{E}_\varepsilon\left[\mathbb{E}_d[\cdot]\right]$.

We argue in the same way than in Section 4 and we use $(\mathbf{A}_3)$ to get

$$\begin{aligned}
\mathbb{E}_{\varepsilon,d}\left[\|s-\tilde{s}\|_n^2\right] &\leqslant C\inf_{m\in\mathcal{M}}\left\{\mathbb{E}_d\left[\|s-s_m\|_n^2 + \frac{\mathrm{Tr}({}^tP_n\pi_mP_n)}{n}\sigma^2\right]\right\} + C'(1+\rho)^2\left(\mathbb{E}_d[\|t-\pi_{F+G}t\|_n^2] + \frac{R}{n}\sigma^2\right) \\
&\leqslant C\inf_{m\in\mathcal{M}}\left\{\mathbb{E}_d[\|s-s_m\|_n^2] + \frac{\dim(S_m)}{n}\rho^2\sigma^2\right\} + C'(1+\rho)^2\left(\mathbb{E}_d[\|t-\pi_{F+G}t\|_n^2] + \frac{R}{n}\sigma^2\right)\ .
\end{aligned}$$

The definition of the norm $\|\cdot\|_n$ implies that, for any $f \in \mathbb{L}^2([0,1],\nu)$,

$$\mathbb{E}_d\left[\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2\right] = \int_0^1 f(x)^2 \nu(dx) \ .$$

Since $s \in \mathcal{H}_\alpha(R)$, it is easy to see that this function lies in a Besov ball. Thus, we can apply Theorem 1 of [9] and we get, for any $m \in \mathcal{M}$,

$$\mathbb{E}_d[\|s - s_m\|_n^2] \leqslant C(\alpha, R)\dim(S_m)^{-2\alpha} \ .$$

Arguing in the same way for the $t_j \in \mathbb{L}^2([0,1],\nu_j)$ and, since $F \perp G$, we obtain

$$\begin{aligned}
\mathbb{E}_d[\|t - \pi_{F+G}t\|_n^2] &\leqslant C(K)\sum_{j=1}^{K} \mathbb{E}_d[\|t^j - \pi_{F+G}t^j\|_n^2] \\
&\leqslant C(K)\sum_{j=1}^{K} \mathbb{E}_d[\|t^j\|_n^2 - \|\pi_F t^j\|_n^2 - \|\pi_G t^j\|_n^2] \\
&\leqslant C(K)\sum_{j=1}^{K} \mathbb{E}_d[\|t^j - \pi_F t^j\|_n^2 - \|t^j - \pi_{E+F} t^j\|_n^2] \\
&\leqslant C(K)\sum_{j=1}^{K} \mathbb{E}_d[\|t^j - \pi_F t^j\|_n^2] \\
&\leqslant C(\alpha, R, K) D_n^{-2\alpha} \leqslant C(\alpha, R, K)\dim(S_m)^{-2\alpha} \ .
\end{aligned}$$

Consequently, for any $m \in \mathcal{M}$, we obtain

$$E_{\varepsilon,d}\left[\|s - \tilde{s}\|_n^2\right] \leqslant C''\left(\dim(S_m)^{-2\alpha} + \frac{\dim(S_m)}{n} + \frac{1}{n}\right) \ .$$

Since $\alpha > \zeta_n$, we can consider some model S_m in $\mathcal{F}^{BM}$ with dimension of order $n^{1/(2\alpha+1)}$ and derive that

$$E_{\varepsilon,d}\left[\|s - \tilde{s}\|_n^2\right] \leqslant C''\left(2n^{-2\alpha/(2\alpha+1)} + \frac{1}{n}\right) \leqslant C_\alpha n^{-2\alpha/(2\alpha+1)} \ .$$

8 Lemmas

This section is devoted to some technical results and their proofs.

Lemma 8.1. *Let $p, q > 0$ be two real numbers such that $2q < p$. For any $\theta > 0$, the following inequality holds*

$$\int_0^\infty \frac{q z^{q-1}}{(\theta + z)^{p/2}} dz \leqslant C(p,q)\theta^{q-p/2}$$

where $C(p,q) = p/(p-2q)$.

Proof. By splitting the integral around θ, we get

$$\begin{aligned}\int_0^\infty \frac{qz^{q-1}}{(\theta+z)^{p/2}}dz &= \int_0^\theta \frac{qz^{q-1}}{(\theta+z)^{p/2}}dz + \int_\theta^\infty \frac{qz^{q-1}}{(\theta+z)^{p/2}}dz \\ &\leqslant \theta^{-p/2}\int_0^\theta qz^{q-1}dz + \int_\theta^\infty qz^{q-1-p/2}dz \\ &\leqslant \left(1+\frac{2q}{p-2q}\right)\theta^{q-p/2}\ .\end{aligned}$$

□

The next lemma is a variant of a lemma due to Laurent and Massart.

Lemma 8.2. *Let $A \in \mathbb{M}_n \setminus \{0\}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ be a standard Gaussian vector of $\mathbb{R}^n$. For any $x > 0$, we have*

$$\mathbb{P}\left(n\|A\varepsilon\|_n^2 \geqslant Tr(A\,{}^tA) + 2\sqrt{\rho(A)^2\,Tr(A\,{}^tA)x} + 2\rho(A)^2x\right) \leqslant e^{-x} \tag{79}$$

and

$$\mathbb{P}\left(n\|A\varepsilon\|_n^2 \leqslant Tr(A\,{}^tA) - 2\sqrt{\rho(A)^2\,Tr(A\,{}^tA)x}\right) \leqslant e^{-x}\ . \tag{80}$$

Proof. It is known that $A\varepsilon$ is a centered Gaussian vector of $\mathbb{R}^n$ of covariance matrix given by the positive symmetric matrix $A\,{}^tA$. Let us denote by $a_1, \ldots, a_n \geqslant 0$ the eigenvalues of the $A\,{}^tA$. Thus, the distribution of $n\|A\varepsilon\|_n^2$ is the same as the one of $\sum_{i=1}^n a_i\varepsilon_i^2$. We have

$$\rho(A)^2 = \max_{i=1,\ldots,n}|a_i| \quad \text{and} \quad \mathrm{Tr}(A\,{}^tA) = \sum_{i=1}^n a_i\ .$$

Because the a_i's are nonnegative,

$$\sum_{i=1}^n a_i^2 \leqslant \rho(A)^2\mathrm{Tr}(A\,{}^tA)$$

and we can apply the Lemma 1 of [22] to obtain the announced inequalities. □

We now introduce some properties that are satisfied by the estimator $\hat{\sigma}^2$ defined in (22).

Lemma 8.3. *In the Gaussian case or under moment condition, the estimator $\hat{\sigma}^2$ satisfies*

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2 + \frac{n\|s-\pi s\|_n^2}{Tr\,({}^tP_n(I_n-\pi)P_n)}\ .$$

Proof. We have the following decomposition

$$\|Y-\pi Y\|_n^2 = \|s-\pi s\|_n^2 + \sigma^2\|(I_n-\pi)P_n\varepsilon\|_n^2 + 2\sigma\langle s-\pi s, P_n\varepsilon\rangle_n\ . \tag{81}$$

The components of ε are independent and centered with unit variance. Thus, taking the expectation on both side, we obtain

$$\mathbb{E}\left[\|Y-\pi Y\|_n^2\right] = \|s-\pi s\|_n^2 + \sigma^2\frac{\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)}{n}\ .$$

□

Lemma 8.4. *Consider the estimator $\hat{\sigma}^2$ defined in the Gaussian case. For any $0 < \delta < 1/2$,*

$$\mathbb{P}\left(\hat{\sigma}^2 \leqslant (1-2\delta)\sigma^2\right) \leqslant C_\delta \exp\left(-\frac{\delta^2\, Tr({}^tP_nP_n)}{16\rho^2(P_n)}\right)$$

where $C_\delta > 1$ only depends on δ.

Proof. Let $a \in V^\perp$ such that $\|a\|_n^2 = 1$, we set

$$u = \begin{cases} (s-\pi s)/\|s-\pi s\|_n & \text{if } s \neq \pi s\ , \\ a & \text{otherwise}\ . \end{cases}$$

We have

$$\begin{aligned} 2\sigma|\langle s-\pi s, P_n\varepsilon\rangle_n| &= 2\sigma|\langle u, P_n\varepsilon\rangle_n| \times \|s-\pi s\|_n \\ &\leqslant \|s-\pi s\|_n^2 + \sigma^2\langle u, P_n\varepsilon\rangle_n^2 \end{aligned}$$

and we deduce from (81)

$$\begin{aligned} \|Y-\pi Y\|_n^2 &\geqslant \sigma^2\|(I_n-\pi)P_n\varepsilon\|_n^2 - \sigma^2\langle u, P_n\varepsilon\rangle_n^2 \\ &= \sigma^2\left(\|P_n\varepsilon\|_n^2 - \left(\|\pi P_n\varepsilon\|_n^2 + \langle u, P_n\varepsilon\rangle_n^2\right)\right) \\ &= \sigma^2\left(\|P_n\varepsilon\|_n^2 - \|\pi' P_n\varepsilon\|_n^2\right) \end{aligned} \tag{82}$$

where π' is the orthogonal projection onto $V \oplus \mathbb{R}u$. Consequently,

$$\begin{aligned} \mathbb{P}\left(\hat{\sigma} \leqslant (1-2\delta)\sigma^2\right) &\leqslant \mathbb{P}\left(n\|P_n\varepsilon\|_n^2 - n\|\pi' P_n\varepsilon\|_n^2 \leqslant (1-2\delta)\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \\ &\leqslant \mathbb{P}\left(n\|P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_nP_n) \leqslant -\delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \\ &\qquad + \mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi P_n) \geqslant \delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \\ &= \mathbb{P}_1 + \mathbb{P}_2\ . \end{aligned} \tag{83}$$

The Inequality (80) and (21) give us the following upperbound for $\mathbb{P}_1$,

$$\mathbb{P}_1 \leqslant \exp\left(-\frac{\delta^2\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)^2}{4\rho^2(P_n)\mathrm{Tr}({}^tP_nP_n)}\right) \leqslant \exp\left(-\frac{\delta^2\mathrm{Tr}({}^tP_nP_n)}{16\rho^2(P_n)}\right)\ . \tag{84}$$

By the properties of the norm ρ, we deduce that

$$\mathrm{Tr}({}^tP_n\pi' P_n) = \mathrm{Tr}({}^tP_n\pi P_n) + \mathrm{Tr}({}^tP_n\pi_u P_n) \leqslant \mathrm{Tr}({}^tP_n\pi P_n) + \rho^2(P_n) \tag{85}$$

where we have defined π_u as the orthogonal projection onto $\mathbb{R}u$. We now apply (79) with $A = \pi' P_n$ to obtain, for any $x > 0$,

$$\begin{aligned} &\mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi P_n) + (1+\delta/2)\rho^2(P_n) + (2+2/\delta)x\right) \\ &\quad\leqslant \mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi' P_n) + (2+2/\delta)x\right) \\ &\quad\leqslant \mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi' P_n) \geqslant 2\sqrt{\mathrm{Tr}({}^tP_n\pi' P_n)x} + 2x\right) \\ &\quad\leqslant \exp\left(-x/\rho^2(\pi' P_n)\right) \\ &\quad\leqslant \exp\left(-x/\rho^2(P_n)\right)\ . \end{aligned}$$

Obviously, this inequality can be extended to $x \in \mathbb{R}$,

$$\mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi P_n) + (1+\delta/2)\rho^2(P_n) + (2+2/\delta)x\right) \leqslant \exp\left(-\frac{x\vee 0}{\rho^2(P_n)}\right) \tag{86}$$

and we take

$$\begin{aligned} x &= \frac{\delta}{2(\delta+1)}\left(\delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n) - \frac{\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right) \\ &= \frac{\delta}{2(\delta+1)}\left(\delta\mathrm{Tr}({}^tP_nP_n) - \frac{3\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right) \\ &\geqslant \frac{\delta}{2(\delta+1)}\left(\frac{\delta\mathrm{Tr}({}^tP_nP_n)}{4} - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right) . \end{aligned}$$

Finally, we get

$$\begin{aligned} \mathbb{P}_2 &\leqslant \exp\left(-\frac{\delta}{2(\delta+1)\rho^2(P_n)}\left(\frac{\delta\mathrm{Tr}({}^tP_nP_n)}{4} - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right)_+\right) \\ &\leqslant \exp\left(-\frac{\delta(\delta+2)}{4(\delta+1)}\left(\frac{\delta\mathrm{Tr}({}^tP_nP_n)}{2(\delta+2)\rho^2(P_n)} - 1\right)_+\right) \\ &= \left\{\exp\left(\frac{\delta(\delta+2)}{4(\delta+1)}\right) \times \exp\left(-\frac{\delta^2\mathrm{Tr}({}^tP_nP_n)}{8(\delta+1)\rho^2(P_n)}\right)\right\} \wedge 1 . \end{aligned} \tag{87}$$

To conclude, we use (84) and (87) in (83). □

Lemma 8.5. *Consider the estimator $\hat{\sigma}^2$ defined under moment condition. For any $0 < \delta < 1/3$, there exists a sequence $(\kappa_{\delta,n})_{n\in\mathbb{N}}$ of positive numbers that tends to a positive constant κ_δ as $Tr({}^tP_nP_n)/\rho^2(P_n)$ tends to infinity, such that*

$$\mathbb{P}\left(\hat{\sigma}^2 \leqslant (1-3\delta)\sigma^2\right) \leqslant C(p,\delta)\kappa_{\delta,n}\tau_p\rho^{(p-2)\vee 2}(P_n)\, Tr({}^tP_nP_n)^{-((p/2-1)\wedge 1)} .$$

Proof. We define the vector $u \in V^\perp$ and the projection matrix π' as we did in the proof of Lemma 8.4. The lowerbound (82) does not depend on the distribution of ε and gives

$$\mathbb{P}\left(\hat{\sigma}^2 \leqslant (1-3\delta)\sigma^2\right) \leqslant \mathbb{P}\left(n\|P_n\varepsilon\|_n^2 - n\|\pi' P_n\varepsilon\|_n^2 \leqslant (1-3\delta)\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) . \tag{88}$$

Since the matrix ${}^tP_nP_n$ is symmetric, we have the following decomposition

$$\begin{aligned} n\|P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_nP_n) &= n\langle {}^tP_nP_n\varepsilon,\varepsilon\rangle_n - \mathrm{Tr}({}^tP_nP_n) \\ &= \sum_{i=1}^n\sum_{j=1}^n({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j - \mathrm{Tr}({}^tP_nP_n) \\ &= \sum_{i=1}^n({}^tP_nP_n)_{ii}(\varepsilon_i^2-1) + 2\sum_{i=1}^n\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j . \end{aligned}$$

Thus, (88) leads to

$$\mathbb{P}\left(\hat{\sigma}^2 \leqslant (1-3\delta)\sigma^2\right) \leqslant \bar{\mathbb{P}}_1 + \bar{\mathbb{P}}_2 + \bar{\mathbb{P}}_3 \tag{89}$$

where we have set

$$\bar{\mathbb{P}}_1 = \mathbb{P}\left(\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1) \leqslant -\delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \ ,$$

$$\bar{\mathbb{P}}_2 = \mathbb{P}\left(2\sum_{i=1}^{n}\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j \leqslant -\delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right)$$

and

$$\bar{\mathbb{P}}_3 = \mathbb{P}\left(n\|\pi' P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi P_n) \geqslant \delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \ .$$

Note that $\bar{\mathbb{P}}_1$ concerns a sum of independent centered random variables. By Markov's inequality and (21), we get

$$\begin{aligned}
\bar{\mathbb{P}}_1 &\leqslant \mathbb{P}\left(\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right| \geqslant \delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \\
&\leqslant \delta^{-p/2}\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)^{-p/2}\mathbb{E}\left[\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2}\right] \\
&\leqslant 2^{p/2}\delta^{-p/2}\mathrm{Tr}({}^tP_nP_n)^{-p/2}\mathbb{E}\left[\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2}\right] \ . \qquad (90)
\end{aligned}$$

If $p \geqslant 4$ then we use the Rosenthal Inequality (see Chapter 2 of [31]) and (57) to obtain

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2}\right] \leqslant C'(p)\tau_p\left(\sum_{i=1}^{n}({}^tP_nP_n)_{ii}^{p/2} + \left(\sum_{i=1}^{n}({}^tP_nP_n)_{ii}^2\right)^{p/4}\right) \ .$$

Since, for any $i \in \{1,\dots,n\}$, $({}^tP_nP_n)_{ii} \leqslant \rho^2(P_n)$, by a convexity argument, we get

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2}\right] \leqslant 2C'(p)\tau_p\rho^{p/2}(P_n)\mathrm{Tr}({}^tP_nP_n)^{p/4} \ .$$

If $2 < p < 4$, we refer to [39] for the following inequality

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2}\right] \leqslant 2\sum_{i=1}^{n}\left|({}^tP_nP_n)_{ii}(\varepsilon_i^2-1)\right|^{p/2} \leqslant C''(p)\tau_p\rho^{p-2}(P_n)\mathrm{Tr}({}^tP_nP_n) \ .$$

In both cases, (90) becomes

$$\bar{\mathbb{P}}_1 \leqslant C(p)\delta^{-p/2}\tau_p\rho^{p/2}(P_n)\mathrm{Tr}({}^tP_nP_n)^{-\beta} \qquad (91)$$

with $\beta = (p/2-1)\wedge p/4$.

Let us now bound $\bar{\mathbb{P}}_2$. By Chebyshev's inequality, we get

$$\begin{aligned}
\bar{\mathbb{P}}_2 &\leqslant \mathbb{P}\left(\left|2\sum_{i=1}^{n}\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j\right| \geqslant \delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)\right) \\
&\leqslant \delta^{-2}\mathrm{Tr}({}^tP_n(I_n-\pi)P_n)^{-2}\mathbb{E}\left[\left(2\sum_{i=1}^{n}\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j\right)^2\right] \\
&\leqslant 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\sum_{i=1}^{n}\sum_{j>i}\sum_{p=1}^{n}\sum_{q>p}({}^tP_nP_n)_{ij}({}^tP_nP_n)_{pq}\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_p\varepsilon_q] \ .
\end{aligned}$$

Note that, by independence between the components of ε, the expectation in the last sum is not null if and only if $i=p$ and $j=q$ (in this case, its value is 1). Thus, we have

$$\begin{aligned}
\bar{\mathbb{P}}_2 &\leqslant 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\sum_{i=1}^{n}\sum_{j>i}({}^tP_nP_n)_{ij}^2 \\
&\leqslant 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\mathrm{Tr}(({}^tP_nP_n)^2) \\
&\leqslant 4\delta^{-2}\rho^2(P_n)\mathrm{Tr}({}^tP_nP_n)^{-1} \ . \qquad (92)
\end{aligned}$$

We finally focus on $\bar{\mathbb{P}}_3$. Recalling (85), we apply Corollary 5.1 of [4] with $\tilde{A} = {}^tP_n\pi'P_n$ to obtain, for any $x>0$,

$$\begin{aligned}
&\mathbb{P}\left(n\|\pi'P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi P_n) + (1+\delta/2)\rho^2(P_n) + (1+2/\delta)x\right) \\
&\quad\leqslant \mathbb{P}\left(n\|\pi'P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi'P_n) + (1+2/\delta)x\right) \\
&\quad\leqslant \mathbb{P}\left(n\|\pi'P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi'P_n) \geqslant 2\sqrt{\mathrm{Tr}({}^tP_n\pi'P_n)x} + x\right) \\
&\quad\leqslant C(p)\tau_p\mathrm{Tr}({}^tP_n\pi'P_n)\rho(\pi'P_n)^{p-2}x^{-p/2} \\
&\quad\leqslant C(p)\tau_p\mathrm{Tr}({}^tP_nP_n)\rho^{p-2}(P_n)x^{-p/2} \ .
\end{aligned}$$

Thus, for any $x\in\mathbb{R}$, we define

$$\psi(x) = \begin{cases} C(p)\tau_p\mathrm{Tr}({}^tP_nP_n)\rho^{p-2}(P_n)x^{-p/2}\wedge 1 & \text{if } x>0 \\ 1 & \text{if } x\leqslant 0 \end{cases}$$

and $\psi(x)$ is an upperbound for

$$\mathbb{P}\left(n\|\pi'P_n\varepsilon\|_n^2 \geqslant (1+\delta/2)\mathrm{Tr}({}^tP_n\pi P_n) + (1+\delta/2)\rho^2(P_n) + (1+2/\delta)x\right) \ .$$

If we take

$$\begin{aligned}
x &= \frac{\delta}{\delta+2}\left(\delta\mathrm{Tr}({}^tP_n(I_n-\pi)P_n) - \frac{\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right) \\
&= \frac{\delta}{\delta+2}\left(\delta\mathrm{Tr}({}^tP_nP_n) - \frac{3\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right) \\
&\geqslant \frac{\delta}{\delta+2}\left(\frac{\delta\mathrm{Tr}({}^tP_nP_n)}{4} - \left(1+\frac{\delta}{2}\right)\rho^2(P_n)\right)
\end{aligned}$$

then we obtain

$$\begin{aligned}\bar{\mathbb{P}}_3 &\leqslant C'(p,\delta)\tau_p \frac{\mathrm{Tr}({}^tP_nP_n)\rho^{p-2}(P_n)}{(\delta\mathrm{Tr}({}^tP_nP_n)/4-(1+\delta/2)\,\rho^2(P_n))_+^{p/2}} \wedge 1 \\ &\leqslant C''(p,\delta)\tau_p \frac{\mathrm{Tr}({}^tP_nP_n)^{1-p/2}\rho^{p-2}(P_n)}{(1-2\,(1+2/\delta)\,\rho^2(P_n)/\mathrm{Tr}({}^tP_nP_n))_+^{p/2}} \wedge 1 \end{aligned} \tag{93}$$

To conclude, we use (91), (92) and (93) in (89). □

References

[1] H. Akaike. Statistical predictor identification. *Annals of the Institute for Statistical Mathematics*, 22:203–217, 1970.

[2] S. Arlot. Choosing a penalty for model selection in heteroscedastic regression. Arxiv preprint arXiv:0812.3141v2, 2010.

[3] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[4] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[5] Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

[6] Y. Baraud, F. Comte, and G. Viennet. Adaptive estimation in autoregression or β-mixing regression via model selection. *Annals of Statistics*, 29(3):839–875, 2001.

[7] L. Birgé and P. Massart. From model selection to adaptive estimation. *Festschrift for Lucien Lecam: Research Papers in Probability and Statistics*, pages 55–87, 1997.

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

[9] L. Birgé and P. Massart. An adaptive compression algorithm in Besov spaces. *Constructive Approximation*, 16:1–36, 2000.

[10] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[11] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[12] L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association*, 80(391):580–619, 1985.

[13] E. Brunel and F. Comte. Adaptive nonparametric regression estimation in presence of right censoring. *Mathematical Methods of Statistics*, 15(3):233–255, 2006.

[14] E. Brunel and F. Comte. *Model selection for additive regression models in the presence of censoring*, chapter 1 in "Mathematical Methods in Survival Analysis, Reliability and Quality of Life", pages 17–31. Wiley, 2008.

[15] A. Buja, T.J. Hastie, and R.J. Tibshirani. Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–555, 1989.

[16] F. Comte and Y. Rozenholc. Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Processes and Their Applications*, 97:111–145, 2002.

[17] X. Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. *Electronic Journal of Statistics*, 2:1345–1372, 2008.

[18] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.

[19] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[20] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

[21] B. Laurent, J.M. Loubes, and C. Marteau. Testing inverse problems: a direct or an indirect problem? *Journal of Statistical Planning and Inference*, 141:1849–1861, 2011.

[22] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[23] W. Leontief. Introduction to a theory of the internal structure of functional relationships. *Econometrica*, 15:361–373, 1947.

[24] O. Linton and J.P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–101, 1995.

[25] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[26] E. Mammen, O. Linton, and J.P. Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27:1443–1490, 1999.

[27] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003.

[28] A.D.R. McQuarrie and C.L. Tsai. *Regression and times series model selection*. River Edge, NJ, 1998.

[29] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[30] J. Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25:186–211, 1997.

[31] V.V. Petrov. *Limit theorems of probability theory: sequences of independent random variables*. Oxford Studies in Probability 4, 1995.

[32] P.D. Ravikumar, H. Liu, J.D. Lafferty, and L.A. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society*, 71:1009–1030, 2009.

[33] S. Robin, F. Rodolphe, and S. Schbath. *DNA, Words and Models*. Cambridge University Press, 2005.

[34] D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *Annals of Statistics*, 22(3):1346–1370, 1994.

[35] H. Scheffé. *The analysis of variance*. Wiley-Interscience, 1959.

[36] E. Severance-Lossin and S. Sperlich. Estimation of derivatives for additive separable models. *Statistics*, 33:241–265, 1999.

[37] C.J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 14(2):590–606, 1985.

[38] D. Tjøstheim and B. Auestad. Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association*, 89:1410–1430, 1994.

[39] B. von Bahr and C.G. Esseen. Inequalities for the rth absolute moment of a sum of random variables $1 \leqslant r \leqslant 2$. *Annals of Mathematical Statistics*, 36:299–303, 1965.