Gaussian linear model selection in a dependent context

Emmanuel Caron, Jérôme Dedecker, Bertrand Michel

Gaussian linear model selection in a dependent context

Emmanuel Caron* Jérôme Dedecker† Bertrand Michel‡

January 28, 2021

Abstract

In this paper, we study the nonparametric linear model, when the error process is a dependent Gaussian process. We focus on the estimation of the mean vector via a model selection approach. We first give the general theoretical form of the penalty function, ensuring that the penalized estimator among a collection of models satisfies an oracle inequality. Then we derive a penalty shape involving the spectral radius of the covariance matrix of the errors, which can be chosen proportional to the dimension when the error process is stationary and short range dependent. However, this penalty can be too rough in some cases, in particular when the error process is long range dependent. In a second part, we focus on the fixed-design regression model assuming that the error process is a stationary Gaussian process. We propose a model selection procedure in order to estimate the mean function via piecewise polynomials on a regular partition, when the error process is either short range dependent, long range dependent or anti-persistent. We present different kinds of penalties, depending on the memory of the process. For each case, an adaptive estimator is built, and the rates of convergence are computed. Thanks to several sets of simulations, we study the performance of these different penalties for all types of errors (short memory, long memory and anti-persistent errors). Finally, we give an application of our method to the well-known Nile data, which clearly shows that the type of dependence of the error process must be taken into account.

Keywords : Nonparametric regression, Model selection, Adaptive estimation, Short memory, Long memory

MSC : 62G05, 62M10, 60G22

Contents

*Emmanuel Caron, Avignon Université, Laboratoire de Mathématiques d'Avignon EA2151, 84000 Avignon, France.
Email: emmanuel.caron-parte@univ-avignon.fr, Website: http://ecaron.perso.math.cnrs.fr/index.html

†Jérôme Dedecker, Université de Paris, Laboratoire MAP5 UMR 8145, 75006 Paris, France.
Email: jerome.dedecker@parisdescartes.fr, Website: http://w3.mi.parisdescartes.fr/~jdedecke/

‡Bertrand Michel, Ecole Centrale de Nantes, Laboratoire de Mathématiques Jean Leray UMR 6629, 44300 Nantes, France.
Email: bertrand.michel@ec-nantes.fr, Website: http://bertrand.michel.perso.math.cnrs.fr

1 Introduction

Let us consider the linear model

$$Y = \mu^* + \varepsilon, \tag{1.1}$$

where Y is the n-dimensional vector of observations, μ^* is an unknown (deterministic) vector to be estimated, and ε is the vector of errors. It is well know that Model (1.1) can serve as a canonical model to express a large class of statistical problems (see [BM01a]). In this paper, we focus on the estimation of the vector μ^* with a model selection approach, in the general framework where the error process ε is a dependent Gaussian random vector, with covariance matrix Σ. Our first goal is to give the theoretical form of the penalty function, depending on Σ, ensuring that the penalized estimator among a collection of models satisfies an oracle inequality.

This model has been widely studied for independent and identically distributed (i.i.d.) errors, in particular by Birgé and Massart in the Gaussian case [BM01a]. Baraud worked in the general i.i.d. case with a deterministic design first [Bar00], then with a random design [Bar02]. Some extensions of these results to a β-mixing framework are presented in [BCV01]. The idea of using a penalty function goes back to the pioneering works of Akaike [Aka73] and Mallows [Mal73]. Later, Birgé and Massart developed a non-asymptotic approach to the selection of penalized models [BM01a], [BM01b], [BM07].

We follow in this paper the strategy developed by Birgé and Massart which is based on a non-asymptotic control of the fluctuations of the empirical contrast.

Let us be more precise here. In order to find a linear subspace that realizes a bias-variance tradeoff, let us introduce a finite collection of models $\{S_m, m \in \mathcal{M}\}$, denoting by d_m the dimension of S_m. Let then $\hat{\mu}_m$ be the least squares estimator of μ^* on S_m. A penalization strategy is used by selecting a model with a criterion of the form

$$\hat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \|Y - \hat{\mu}_m\|_n^2 + \operatorname{pen}(m) \right\},$$

where $\|\cdot\|_n$ denotes the (normalized) euclidean norm in $\mathbb{R}^n$, and $\operatorname{pen} : \mathcal{M} \to \mathbb{R}^+$ is a penalty function defined on the family of models (as usual, $\operatorname{argmin}_{m\in\mathcal{M}}$ defines a set of points in $\mathcal{M}$ and the notation $\in$ means that $\hat{m}$ can be any point of that set). Following the Birgé and Massart approach, we derive a penalty function which provides an oracle inequality for the model selection procedure in the dependent Gaussian framework.

In Section 2, a general penalty shape is presented. The main term is the quantity $\operatorname{tr}(\operatorname{Proj}_{S_m}\Sigma)$ (tr denoting the trace and $\operatorname{Proj}_{S_m}$ denoting the matrix of the projection on S_m on the canonical basis) which plays the same role as the term $\operatorname{Var}(\varepsilon_1)\, d_m$ in the results of Birgé and Massart for i.i.d. Gaussian errors. Similar penalties have already been introduced by Gendre [Gen14] in the context of model selection for additive regression. However Gendre [Gen14] is not interested in the same questions as us: he is concerned with additive regression whereas our objective is to study the Gaussian regression with dependent errors. In the same way as for us, the analysis of [Gen14] is based on a general Gaussian model selection, but it appears that for our concern, the general penalty form we provide is more appropriate than that provided by [Gen14]. In addition, the assumptions of [Gen14] do not apply to the context of long range dependent or anti-persistent errors.

Note that the trace $\operatorname{tr}\left(\operatorname{Proj}_{S_m}\Sigma\right)$ is bounded by $d_m\rho(\Sigma)$, where $\rho(\Sigma)$ is the spectral radius of the covariance matrix. Hence, neglecting some residual terms (see Section 2), the following penalty can be used: for any $K > 1$,

$$\operatorname{pen}(m) \geq K\frac{\rho(\Sigma)d_m}{n}. \tag{1.2}$$

For instance, if we suppose that the error process is a short memory stationary process with bounded spectral density, then the spectral radius is bounded, and this penalty shape is very close to the i.i.d. case up to a constant. The penalty can still be chosen proportional to the dimension, as in the i.i.d. case, but the usual variance term is replaced by the spectral radius of the covariance matrix.

However, the penalty (1.2) may be too rough in some cases, in particular if the error process is long range dependent. To see how to handle this case in a concrete situation, we study in Sections 3 and 4 the fixed-design regression model

$$Y_i = f^*\left(\frac{i}{n}\right) + \varepsilon_i\,, \tag{1.3}$$

where $(\varepsilon_i)_{i\geq 1}$ is a stationary Gaussian process. By standard arguments, this model can be written as a special case of the generic Model (1.1) (see the beginning of Section 3).

Note that Model (1.3) has been widely studied in the literature (with possibly non Gaussian errors) via kernel or wavelets methods.

For kernel estimators, let us first quote the paper by Hall and Hart [HH90], who considered a particular class of Gaussian errors. The authors showed in particular that, for a twice differentiable function f^*, the rate is the same as in the i.i.d. case if and only if $\sum_{k>0}|\mathrm{Cov}(\varepsilon_1,\varepsilon_k)|<\infty$, and they gave minimax rates in the long range dependent case. Let us also cite the papers by Csörgő and Mielniczuk [CM95a], [CM95b], [CM95c] (long memory is considered in [CM95b] and [CM95c]), Tran et al [TRYTV96] (short memory case), and Robinson [Rob97]. Robinson's article provides very general results for short range and long range dependent processes, and rates of convergence for anti-persistent errors (also called negatively correlated errors) can be derived from his Lemma 3. Local polynomial fitting with long memory, short memory and anti-persistent errors is considered by Beran and Feng [BF02]. Note that none of these articles adresses the issues of adaptive estimation or data-driven bandwidth selection.

For wavelets type estimators, let us first quote the paper by Wang [Wan96], who gave minimax results in the long range dependent case, when the function f^* belongs to a Besov class. Let us also cite the papers by Johnstone and Silverman [JS97], Johnstone [Joh99], and more recently Li and Xiao [LX07] and Beran and Shumeyko [BS12]. These four papers addressed the issue of a data-driven choice of the threshold. Theorem 1 in [Joh99] gave a very precise minimax result (up to constants), but for an asymptotic model which is a bit different from (1.3) (see the discussion at the end of the paper [Joh99]). By adapting the block thresholding method described in Hall et al [HKP99] to the long memory case, Li and Xiao [LX07] showed that the block thresholded wavelets estimators are adaptive and minimax for a large class of functions.

In Sections 3 and 4 of the present paper, we propose a model selection procedure to estimate f^* via piecewise polynomials on a regular partition of size m. The choice of piecewise polynomials is very natural here, since the function f^* is supported on $[0,1]$, and such estimators do not show bad behaviors near the boundary. We show that

- For short memory error processes (i.e. when $\rho(\Sigma)$ is uniformly bounded) the penalty is of the form

  $$\mathrm{pen}(m) = K\frac{m}{n}$$

  (for some constant $K>0$ to be calibrated), the penalized estimator is adaptive with respect to the unknown regularity of the function f^*, and yields the same rates of convergence as in the i.i.d setting.

- For long memory processes, that is when the auto-covariances $\gamma_\varepsilon(k)$ of the error process are such that

  $$|\gamma_\varepsilon(k)| \leq \kappa k^{-\gamma}, \qquad \text{for some } \kappa>0 \text{ and } \gamma\in(0,1),$$

  the penalty is a concave function of (m/n)

  $$\mathrm{pen}(m) = K\left(\frac{m}{n}\right)^{\gamma}$$

  (for some constant $K>0$ to be calibrated), the penalized estimator is adaptive with respect to the unknown regularity of the function f^*, and yields the same minimax rates of convergence as in [Wan96].

- For anti-persistent errors such that

  $$\mathrm{Var}(\varepsilon_1+\cdots+\varepsilon_n) \leq \kappa n^{2-\gamma}, \qquad \text{for some } \kappa>0 \text{ and } \gamma\in(1,2),$$

  and in the case of regressograms (piecewise polynomials of degree 0), the penalty has the form

  $$\mathrm{pen}(m) = K\left(\frac{m^\gamma}{n^\gamma} + \frac{\log(m)}{n}\right)$$

(for some constant $K > 0$ to be calibrated). The main part of the penalty is then a convex function of (m/n). The penalized estimator is adaptive with respect to the unknown regularity of the function f^*, and yields faster rates of convergence than in the i.i.d setting. Note that similar rates can also be deduced from Lemma 3 in [Rob97].

In Section 4, we simulate different kind of short memory processes (a Gaussian ARMA(2,1) process, two non Gaussian β-mixing Markov chains), of long memory processes (a fractional Gaussian noise with Hurst index in (1/2,1), and a non Gaussian β-mixing Markov chain), and an anti-persistent process (a fractional Gaussian noise with Hurst index in (0, 1/2)). For regressograms on a regular partition of size m, we investigate different kind of penalties: the usual penalty proportional to m/n, a penalty proportional to $(m/n)^\gamma$ in the case of long range dependent or anti-persistent errors, and some penalties for which γ is estimated via an estimator of the Hurst index based on the Y_i's or on the residuals. Finally, an important message of this paper is that the slope heuristics [BM07] can be adapted to calibrate penalties in the context of regression with dependent errors.

In Section 5, we give an application of our method to the well known Nile data, and we continue the discussion started in Robinson's article [Rob97]. In Section 6, we discuss other possible applications of the generals results of Section 2. Finally, Section 7 is devoted to the proofs of the results of Sections 2 and 3.

To conclude this introduction, let us make some additional remarks.

To give a concrete example of application of the general result of Section 2, we have chosen to present the fixed-design regression model (1.3) with all the details. But, as already mentioned in the first paragraph of this introduction, Model (1.1) can be used in many other situations. For instance, let us consider the case of random design regression $Y_i = f^*(X_i) + \varepsilon_i$, where (X_i) and (ε_i) are two independent sequences of stationary random variables, and (ε_i) is a Gaussian sequence. Then, conditioning on the design, our general result still applies. For instance, following closely the proof in Baraud's paper ([Bar02], Gaussian case), we see that, under the same assumptions on the distribution of X and on the collection of models, his main result still holds provided $\rho(\Sigma)$ is uniformly bounded (short memory case). However, the case where the ε_i's are long range dependent is not so clear, a careful study being needed to control the term $\mathrm{tr}(\mathrm{Proj}_{S_m}\,\Sigma)$ (see the Conjecture in Section (6): Discussion).

Our second remark concerns the Gaussian assumption on the error, which is of course quite restrictive. Our guess is that some of our results can be extended to other (non Gaussian) sequences. This is the reason why we consider some short range dependent and long range dependent Harris recurrent Markov chains in the simulation section (Section 4). Based on these experiments, it seems indeed that our method is quite robust, but even for this simple case the theoretical part has to be written properly. As usual for model selection in a Gaussian context, the central tool is Cirel'son-Ibragimov-Sudakov concentration inequality [CIS76] (see Subsection 7.2). Hence, an important problem is to know if some appropriate concentration inequalities can be stated in other dependent contexts that include long range dependent processes.

There are many ways to define long range dependence: through the behavior of autocovariances or of the spectral density, by considering the behavior of the variance of partial sums, via limit theorems for partial sums or for the empirical distribution function. One of these definitions is given in Chapter 3 of the monograph [**?**] (see Definition 3.1.2 there), which contains also many statistical results for long range dependent processes. Among the processes that are known to exhibit long range dependence, an important class is the class of linear processes $\varepsilon_k = \sum_{i=0}^{\infty} a_i \eta_{k-i}$, where $(\eta_i)_{i\in\mathbb{Z}}$ is a sequence of i.i.d. random variables with mean zero and finite variance, and $(a_i)_{i\geq 0}$ is a sequence of real numbers in ℓ^2. These processes can be short range dependent, long range dependent or anti persistent according to the behavior of the a_i's (for instance if the a_i's are summable and $\sum_{i\geq 0} a_i \neq 0$, then the process is short range dependent): a complete description is given in Chapter 3 of [**?**]. Other classes of long memory processes are the squared LARCH processes (introduced by Robinson [**?**], see also [**?**]), linear processes with infinite variance innovation (see Samorodnisky and Taqqu [**?**], and also [**?**]), bounded functions of stationary Markov chains (see for instance [DGM18]), functions of Gaussian processes (see for instance [**?**]) and even some classes of dynamical systems (see [**?**] and [DGM18]). For all these non Gaussian processes, it would be interesting to know if some parts of our results are still valid.

Our last remark concerns the data-based justification of the existence of long memory phenomena. Situations where the error process (ε_k) in Model (1.3) is long range dependent often occur when considering financial or climatology time series. For instance the annual series of winter means of the NAO index (North Atlantic Oscillation index) exhibits long range dependence (see Stephenson et al. [**?**]) and also an increasing trend for the last decade (which can possibly be explained by global warming). Concerning financial time series, we refer to the paper by Pesee [**?**] where daily exchange rate data are studied. The classical monograph by Beran [Ber94] contains also many examples of long memory processes, such as: Northern hemisphere monthly temperatures, ethernet traffic data from a local area network, annual data for Nile river minima

(see Section 5 for more details), weight measurements (deviation from 1kg) from the National Bureau of Standards, ...

2 A Gaussian linear model selection theorem in a dependent context

2.1 General setting

Recall the equation of the Gaussian linear model (1.1)

$$Y = \mu^* + \varepsilon,$$

where the mean vector μ^* belongs to $\mathbb{R}^n$ and where the error vector ε is a Gaussian random vector. We consider the general setting where the components of Y are not necessarily independent

$$\varepsilon \sim \mathcal{N}_n(0, \Sigma).$$

The covariance matrix Σ is a $n \times n$ semidefinite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. We also introduce the spectral radius of Σ

$$\rho(\Sigma) = \max_{1 \leq i \leq n} \lambda_i = \lambda_1.$$

The aim is to estimate the unknown vector μ^* from the observation Y. One standard strategy is to constrain the estimator to belong to a given linear subspace S of $\mathbb{R}^n$. Let $\|\cdot\|_n$ denotes the (normalized) euclidean norm in $\mathbb{R}^n$

$$\|x\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

The least squares contrast is defined for $x \in \mathbb{R}^n$ by

$$\gamma_n(x) = \|Y - x\|_n^2,$$

and the minimizer of γ_n over S is the orthogonal projection of Y on S

$$\mathrm{Proj}_S(Y) = \mathrm{argmin}_{x \in S}\, \gamma_n(x).$$

With a slight abuse of notation, we shall write Proj_S for the projection operator on S and for its matrix on the canonical basis. The ℓ^2 risk of an estimator $\hat{\mu}$ is defined by

$$R(\hat{\mu}) = \mathbb{E}\left[\|\hat{\mu} - \mu^*\|_n^2\right],$$

where the expectation is under the distribution of Y. Using Pythagoras equality in $\mathbb{R}^n$ together with (1.1), we find that the risk of $\mathrm{Proj}_S(Y)$ satisfies the following bias-variance decomposition

$$\mathbb{E}\left[\|\mu^* - \mathrm{Proj}_S(Y)\|_n^2\right] = \|(\mathrm{Id} - \mathrm{Proj}_S)\mu^*\|_n^2 + \mathbb{E}\left[\|\mathrm{Proj}_S(\varepsilon)\|_n^2\right].$$

The squared bias $\|(\mathrm{Id} - \mathrm{Proj}_S)\mu^*\|_n^2$ is small for large enough linear subspace S. It can be easily checked that the variance term is equal to $\mathbb{E}\left[\|\mathrm{Proj}_S(\varepsilon)\|_n^2\right] = \frac{1}{n}\,\mathrm{tr}\,(\mathrm{Proj}_S\,\Sigma)$, see the proof of Theorem 2.1. As in the i.i.d. case, the variance term tends to increase with the dimension of S.

In order to find a linear subspace that realizes a bias-variance tradeoff, we introduce a finite collection of linear subspaces $\{S_m, m \in \mathcal{M}\}$ that we call *models*, and we denote by d_m the dimension of S_m. For $m \in \mathcal{M}$, we denote by $\hat{\mu}_m$ the least squares estimator $\mathrm{Proj}_{S_m}(Y)$ of μ^* on S_m. We also introduce the oracle model m_0, that is the model that provides the least squares estimator with minimum risk

$$m_0 \in \mathrm{argmin}_{m \in \mathcal{M}}\{R(\hat{\mu}_m)\}.$$

Now the aim is to select a model in the collection such that the risk of the selected estimator is as close as possible to the oracle model.

The true risk $R(\hat{\mu}_m)$ of $\hat{\mu}_m$ being unknown in practice, we introduce the empirical risk

$$\widehat{R}(\hat{\mu}_m) = \|Y - \hat{\mu}_m\|_n^2.$$

Obviously this criterion can not be used to select a model in the collection because of the overfitting effect. We follow a penalization strategy [Aka73, Mal73, BM01a, Mas07] by selecting a model with a criterion of the form

$$\hat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \|Y - \hat{\mu}_m\|_n^2 + \operatorname{pen}(m) \right\}, \tag{2.1}$$

where $\operatorname{pen} : \mathcal{M} \to \mathbb{R}^+$ is a penalty function defined on the family of models. In this paper we perform a non asymptotic analysis of the risk of the selected estimator $\hat{\mu}_{\hat{m}}$. By this way we derive a penalty function which provides an oracle inequality for the model selection procedure, in the dependent Gaussian context.

2.2 A general Gaussian model selection result

Let $\pi = \{\pi_m, m \in \mathcal{M}\}$ be a probability measure defined on $\mathcal{M} : \sum_{m\in\mathcal{M}} \pi_m = 1$. We first give a general shape for the penalty function and the corresponding oracle inequality.

Theorem 2.1. *Let $K > 1$, and let $\operatorname{pen} : \mathcal{M} \to \mathbb{R}^+$ be a penalty function such that: for any $m \in \mathcal{M}$,*

$$\operatorname{pen}(m) \geq \frac{K}{n} \left(\sqrt{\operatorname{tr}\left(\operatorname{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2. \tag{2.2}$$

Then there exists a constant $C > 1$ which only depends on K such that the estimator $\hat{\mu}_{\hat{m}}$ selected by the criterion (2.1) satisfies

$$\mathbb{E}\left[\|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2\right] \leq C \left(\inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[\|\mu^* - \hat{\mu}_m\|_n^2\right] + \operatorname{pen}(m) \right\} + \frac{\rho(\Sigma)}{n} \right). \tag{2.3}$$

The main term in the penalty shape (2.2) is the trace term $\operatorname{tr}\left(\operatorname{Proj}_{S_m} \Sigma\right)$. This quantity plays the same role as the term $\operatorname{Var}(\varepsilon_1) d_m$ in the results of Birgé and Massart for independent Gaussian errors [BM01a, Mas07]. Of course, this penalty can only be calculated if the matrix Σ is completely known. However we will see that, in certain cases, we can consider effective strategies to circumvent this issue (see Sections 3 and 4).

We can propose penalty shapes from the upper bounds

$$\operatorname{tr}\left(\operatorname{Proj}_{S_m} \Sigma\right) \leq \sum_{i=1}^{d_m} \lambda_i \leq d_m \rho(\Sigma). \tag{2.4}$$

A proof for the first inequality is given in Subsection 7.5. Actually, with a minor modification of the proof of Theorem 2.1 (this modification being necessary only to get a better constant in front of $\sqrt{\rho(\Sigma)}$), it can be checked that the risk bound (2.3) is still valid when replacing the lower bound in (2.2) by

$$\operatorname{pen}(m) \geq \frac{K}{n} \left(\sqrt{\sum_{i=1}^{d_m} \lambda_i} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2,$$

or by

$$\operatorname{pen}(m) \geq K \frac{\rho(\Sigma)}{n} \left(\sqrt{d_m} + \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2, \tag{2.5}$$

for any $K > 1$.

If the sequence $(\varepsilon_i)_{i\geq 1}$ is a stationary and short memory Gaussian process, then the spectral radius is bounded (see Remark 3.1 below) and the penalty shape (2.5) is completely in line with the case of independent Gaussian errors [BM01a, Mas07], the usual variance term $\operatorname{Var}(\varepsilon_1)$ being replaced by the spectral radius $\rho(\Sigma)$.

The three penalty shapes given above depend on the probability π. If the collection of model is not too rich (see for instance [BM01a, Mas07] or Chapter 2 in [Gir14]), it might be chosen in such a way that

$$\rho(\Sigma) \log\left(\frac{1}{\pi_m}\right)$$

is smaller or of the same order as the main terms $\operatorname{tr}\left(\operatorname{Proj}_{S_m} \Sigma\right)$, $\sum_{i=1}^{d_m} \lambda_i$ or $d_m \rho(\Sigma)$. To sum up, if the spectral radius is bounded and if the collection of models is not too rich, we see that the penalty can be chosen proportional to the dimension d_m, as in the independent case.

It is tempting to keep the penalty shape (2.5) as a general penalty shape for Gaussian linear model selection with dependent errors. However, as we will see later in the paper, this penalty shape is too rough in some cases. For instance, it cannot lead to minimax rates of convergence for non parametric regression with long range dependent errors (see Subsection 3.2).

At this point, it should be clearly quoted that a penalty similar to (2.2) has been given in the paper [Gen14]. The main difference is that, in the inequality similar to (2.3) proved in [Gen14] (Inequality (2.2) of Theorem 2.1 in [Gen14]), the residual term is $\rho(\Sigma)R_n/n$ instead of $\rho(\Sigma)/n$. For the questions he has in mind (which are not directly related to time series), Gendre is able to effectively control this additional term R_n. But it does not seem easy to handle for long range dependent errors or anti-persistent errors, which are precisely the kind of error processes that we want to study in the present paper.

3 Non parametric regression with Gaussian dependent errors

In this section we study the fixed design regression problem with dependent Gaussian errors. Let f^* be a function in $\mathbb{L}^\infty([0,1])$, and recall the equation of model (1.3)

$$Y_i = f^*\left(\frac{i}{n}\right) + \varepsilon_i, \quad i \in \{1, \dots n\},$$

where $(\varepsilon_1, \dots, \varepsilon_n) \sim \mathcal{N}_n(0, \Sigma_n)$. The aim is to estimate f^* based on the observations $Y_1, \dots, Y_n$.

By considering the application

$$f \in \mathbb{L}^\infty([0,1]) \mapsto I(f) = (f(1/n), \dots, f(1)) \in \mathbb{R}^n,$$

we can easily associate a linear subspace of $\mathbb{R}^n$ to any linear subspace of $\mathbb{L}^\infty([0,1])$. Slightly abusing the notation, we identify the function f to the vector $I(f)$, and we write

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(i/n) = \|I(f)\|_n^2, \text{ for } f \in \mathbb{L}^\infty([0,1]).$$

For F a finite linear subspace of $\mathbb{L}^\infty([0,1])$, we define the least squares estimator $\hat{f}$ of f^* on F as

$$\hat{f} = \operatorname{argmin}_{f\in F} \|Y - f\|_n^2, \quad \text{where} \quad \|Y - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(i/n) - Y_i)^2.$$

We shall only consider here the linear spaces S_m of $\mathbb{R}^n$ induced by the linear space F_m of $\mathbb{L}^\infty([0,1])$ generated by the family of piecewise polynomials of degree at most r ($r \in \mathbb{N}$) on the regular partition of size m of the interval $[0,1]$. Obviously, the linear space S_m has dimension $d_m = (r+1)m$; the case $r = 0$ corresponds to the regular regressogram of size m.

We denote by $\hat{f}_m$ the least square estimator of f^* on F_m.

We shall always consider some weights π_m of order m^{-2} (suitably normalized in such a way that $\sum_{m=1}^n \pi_m = 1$). For such weights, the terms involving π_m in the general penalty (2.2) is of order $\rho(\Sigma)\log(m)$; in the applications given below, it will be negligible with respect to the main term $\operatorname{tr}\left(\operatorname{Proj}_{S_m}\Sigma\right)$.

3.1 The case of short range dependent sequences

In this subsection, we assume that the error process $(\varepsilon_i)_{i\geq 1}$ is stationary and short-range dependent. By short range dependent, we mean that

$$\rho_\varepsilon = \sup_{n\in\mathbb{N}^*} \rho(\Sigma_n) < \infty. \tag{3.1}$$

From (3.1), we immediately see that $\operatorname{Var}(\varepsilon_1 + \dots + \varepsilon_n) \leq n\rho_\epsilon$; the fact that the variance of partial sums does not grow faster than n is always required for short-range dependence. Our definition is however a bit less restrictive than the usual definitions based on the spectral density or on the summability of the auto-covariances $\gamma_\varepsilon(k) = \operatorname{Cov}(\varepsilon_0, \varepsilon_k)$, as explained in the following remark.

Remark 3.1. *By definition of the spectral radius and of the spectral density g_ε of $(\varepsilon_i)_{i\geq 1}$, we have*

$$\rho(\Sigma_n) = \sup_{n\|x\|_n^2=1} x^t\Sigma_n x = \sup_{n\|x\|_n^2=1} \operatorname{Var}\left(\sum_{k=1}^n x_k\varepsilon_k\right)$$
$$= \sup_{n\|x\|_n^2=1} \int_{-\pi}^{\pi} \left|\sum_{k=1}^n x_k e^{ikx}\right|^2 g_\varepsilon(x)dx\,.$$

It follows that

$$\rho_\varepsilon \le 2\pi \|g_\varepsilon\|_\infty \le \gamma_\varepsilon(0) + 2\sum_{k=1}^{\infty} |\gamma_\varepsilon(k)|\,,$$

Hence Condition (3.1) *is implied by the boundedness of* g_ε *(and therefore also by the summability of the* $\gamma_\varepsilon(k)$*'s).*

Now, if (3.1) holds, the model selection procedure is exactly the same as in the i.i.d. framework, by replacing the variance of the errors by the spectral radius in the penalty. More precisely, we obtain a penalty of the form

$$\mathrm{pen}(m) = K\rho_\varepsilon \frac{m}{n}\,,$$

for some positive constant K depending on the degree r. We now select a model in $\mathcal{M}_n$ according to the criterion (2.1), which can be rewritten as

$$\hat{m} \in \mathrm{argmin}_{m\in\{1,\dots,n\}} \left\{ \left\| Y - \hat{f}_m \right\|_n^2 + \mathrm{pen}(m) \right\}. \tag{3.2}$$

Following [Bar00], we derive rates of convergence when f^* belongs to some Besov spaces $\mathcal{B}_{\alpha,\ell,\infty}$ for $\ell^{-1} < \alpha < r+1$ and $\ell \ge 2$ (see [DL93] for the definition of Besov spaces). In short, the approximation term in the risk decomposition of $\hat{f}_m$ satisfies (see Sections 4 and 7.4 in [Bar00])

$$\inf_{g\in F_m} \|f^* - g\|_n^2 \le C(\alpha, r)|f^*|_{\alpha,\ell}^2 \left(m^{-2\alpha} + n^{-2\alpha+2/\ell} \right)\,, \tag{3.3}$$

where $|\cdot|_{\alpha,\ell}$ is the usual norm on $\mathcal{B}_{\alpha,\ell,\infty}$. Balancing the variance term and the approximation terms exactly as in case of i.i.d errors, we end up with the same rate of convergence as in the i.i.d. case.

Corollary 3.1. *Let* (ℓ, α) *be such that* $\alpha \in (0, r+1)$ *and* $\ell \ge \max(2, (2\alpha+1)/(2\alpha^2))$*. For a stationary Gaussian process satisfying* (3.1)*, and for the estimator* $\hat{f}_{\hat{m}}$ *selected according to the penalized criterion procedure* (3.2)*,*

$$\sup_{|f^*|_{\alpha,\ell}\le L} \mathbb{E} \left\| f^* - \hat{f}_{\hat{m}} \right\|_n^2 \le C n^{-\frac{2\alpha}{2\alpha+1}},$$

where C *depends on* ρ_ε*,* K*,* α*,* ℓ *and* L*.*

This upper bound is known to be the minimax rate of convergence for the estimation of f^* in the i.i.d. case. This is satisfactory since a sequence of i.i.d. Gaussian random variables is of course short-range dependent.

As for the Gaussian i.i.d case, the penalty is defined up to a multiplicative constant K. The spectral radius is unknown, as is the variance of the errors in the standard i.i.d. setting. In practice, the penalty is chosen proportional to the model dimension m and calibrated according to the slope heuristic method introduced by Birgé et Massart [BM01b], see Section 4.2 further.

3.2 The case of long range dependent sequences

In this subsection, we assume that the error process $(\varepsilon_i)_{i\ge1}$ is strictly stationary, but we do not assume that (3.1) holds. Instead, we assume that

$$|\gamma_\varepsilon(k)| \le \kappa k^{-\gamma}, \quad \text{for some } \kappa > 0 \text{ and } \gamma \in (0,1), \tag{3.4}$$

where $\gamma_\varepsilon(k)$ is the auto-covariance $\gamma_\varepsilon(k) = \mathrm{Cov}(\varepsilon_0, \varepsilon_k)$. Of course, (3.4) is only an upper bound, so that it may happen that $\sum_{k>0} |\gamma_\varepsilon(k)| < \infty$; in such a case (3.1) holds and the process is short range dependent. But the interesting case is of course when $|\gamma_\varepsilon(k)|$ is exactly of order $k^{-\gamma}$, so that $\sum_{k>0} |\gamma_\varepsilon(k)| = \infty$. This is what we mean here by long range dependent.

To control the main term of the penalty, we shall prove the following lemma

Lemma 3.1. *Let* S_m *be the linear space of* $\mathbb{R}^n$ *induced by the family of piecewise polynomials of degree at most* r *on the regular partition of size* m *of the interval* $[0,1]$*. If* (3.4) *holds, then*

$$\mathrm{tr}\left(\mathrm{Proj}_{S_m}\,\Sigma\right) \le C m^\gamma n^{1-\gamma}\,,$$

where C *depends on* κ, γ *and* r*.*

Moreover, by the classical Gerschgorin theorem [**?**], we easily see that

$$\rho(\Sigma_n) \leq B n^{1-\gamma}\,,$$

where B depends on κ and γ. Combining this last bound with Lemma 3.1, we infer from (2.2) that one can choose a penalty of the form

$$\mathrm{pen}(m) = K\frac{m^\gamma}{n^\gamma}\,,$$

for some positive constant K depending on κ, γ and r.

Now, since the bias term (3.3) is still valid for any function f^* in the Besov space $\mathcal{B}_{\alpha,\ell,\infty}$ (with $\ell^{-1} < \alpha < r+1$ and $\ell \geq 2$), we can proceed as in Section 3.1 to get the rate of convergence of the estimator $\hat{f}_{\hat{m}}$. The difference is that the bias-variance problem consists of balancing two terms of order

$$\frac{1}{m^{2\alpha}}\ \text{(bias)} \quad \text{and} \quad \frac{m^\gamma}{n^\gamma}\ \text{(variance)}.$$

This leads to the following corollary.

Corollary 3.2. *Let (ℓ,α) be such that $\alpha \in (0, r+1)$ and $\ell \geq \max(2,(2\alpha+\gamma)/(2\alpha^2))$. For a stationary Gaussian process satisfying (3.4), and for the estimator $\hat{f}_{\hat{m}}$ selected according to the penalized criterion procedure (3.2),*

$$\sup_{|f^*|_{\alpha,\ell}\leq L} \mathbb{E}\left\|f^* - \hat{f}_{\hat{m}}\right\|_n^2 \leq C n^{-\frac{2\alpha\gamma}{2\alpha+\gamma}},$$

where C depends on γ, K, α, ℓ and L.

This rate is satisfactory, since it corresponds to the minimax rates described in the same setting by Wang [Wan96] when $\gamma_\varepsilon(k)$ is exactly of order $k^{-\gamma}$. Note however that the minimax rate in [Wan96] is written for the usual $\mathbb{L}^2([0,1])$-norm.

Let us make some additional comments: if the exponent γ is known, then the slope heuristic can still be used to calibrate the other constants in the penalty term. We shall see that it works pretty well in the simulation section and we will also investigate the calibration of γ for the more general and difficult framework where the exponent γ is unknown.

Remark 3.2. *One can also give an upper bound for $\mathrm{tr}\left(\mathrm{Proj}_{S_m}\Sigma\right)$ in the case where $\gamma_\varepsilon(k) \sim k^{-\gamma}L(k)$, where $\gamma \in (0,1)$ and L is a slowly varying function. From inequality (7.7) of the proof of Lemma 3.1 and the properties of slowly varying functions (applying for instance Proposition 2.2.1 in [**?**]), we obtain*

$$\mathrm{tr}\left(\mathrm{Proj}_{S_m}\Sigma\right) \leq C m^\gamma n^{1-\gamma} L(n/m)$$

for some positive constant C. Using a penalty of the form

$$\mathrm{pen}(m) = K\frac{m^\gamma}{n^\gamma}L(n/m)\,,$$

and following the computations leading to the rate of Corollary 3.2, we end up with the rate

$$\sup_{|f^*|_{\alpha,\ell}\leq L} \mathbb{E}\left\|f^* - \hat{f}_{\hat{m}}\right\|_n^2 \leq C n^{-\frac{2\alpha\gamma}{2\alpha+\gamma}} L\left(n^{\frac{2\alpha}{2\alpha+\gamma}}\right).$$

3.3 Regular regressograms and anti-persistent errors

We now assume that the sequence $(\varepsilon_i)_{i\geq 1}$ is stationary and anti-persistent in the following sense: there exists a parameter $\gamma \in (1,2)$ and a positive constant κ such that Condition (3.1) holds and

$$\mathrm{Var}\left(\sum_{k=1}^n \varepsilon_k\right) \leq \kappa n^{2-\gamma}\,. \tag{3.5}$$

For instance, Conditions (3.1) and (3.5) hold if $(\varepsilon_i)_{i\geq 1}$ is a fractional Gaussian noise with Hurst index $H \in (0,1/2)$ (see Section 4.2 for the definition of the Hurst index). In that case, $\gamma = 2-2H$. The term anti-persistent is borrowed from this particular case.

In this subsection, we only consider the case of regular regressograms, which corresponds to estimators via piecewise polynomials of degree 0 on a regular partition of $[0,1]$.

To control the main term of the penalty, we shall prove the following lemma.

Lemma 3.2. *Let S_m be the linear space of $\mathbb{R}^n$ induced by the family of indicators of intervals on the regular partition of size m of the interval $[0,1]$. If Conditions (3.1) and (3.5) hold, then*

$$\operatorname{tr}\left(\operatorname{Proj}_{S_m}\Sigma\right) \leq Cm^{\gamma}n^{1-\gamma},$$

where C depends on κ and γ.

We infer from (2.2) that one can choose a penalty of the form

$$\operatorname{pen}(m) = K\left(\frac{m^{\gamma}}{n^{\gamma}} + \frac{\log(m)}{n}\right),$$

for some positive constant K depending on κ, γ and ρ_{ε} (recall that ρ_{ε} is the constant appearing in (3.1)).

Now, since the bias term (3.3) is still valid for any function f^* in the Besov space $\mathcal{B}_{\alpha,\ell,\infty}$ (with $\ell^{-1} < \alpha < 1$ and $\ell \geq 2$), we can proceed as in Section 3.1 to get the rate of convergence of the estimator $\hat{f}_{\hat{m}}$. This leads to the following corollary.

Corollary 3.3. *Let (ℓ,α) be such that $\alpha \in (0,1)$ and $\ell \geq \max(2,(2\alpha+\gamma)/(2\alpha^2))$. For a stationary Gaussian process satisfying Conditions (3.1) and (3.5), and for the estimator $\hat{f}_{\hat{m}}$ selected according to the penalized criterion procedure (3.2),*

$$\sup_{|f^*|_{\alpha,\ell}\leq L} \mathbb{E}\left\|f^* - \hat{f}_{\hat{m}}\right\|_n^2 \leq Cn^{-\frac{2\alpha\gamma}{2\alpha+\gamma}},$$

where C depends on γ, K, α, ℓ and L.

It is interesting to notice that, for a regularity $\alpha < 1$, the rate of convergence given in Corollary 3.3 is faster than in the case where the sequence $(\varepsilon_i)_{i\geq 1}$ is i.i.d. It would of course be interesting to know if this rate is minimax (to our knowledge, this has not yet been proven).

4 Numeric experiments

4.1 Slope heuristics

For the results given in the previous sections, the penalty functions are known, in the best case, up to a multiplicative constant. The aim of the slope heuristics method proposed by Birgé and Massart [BM07] is precisely to calibrate a penalty function for model selection purposes. See [BMM12] and [Arl19] for a general presentation of the method. This method has shown very good performances and comes with mathematical guarantees for non parametric Gaussian regression with i.i.d. error terms, see [BM07, Arl19] and references therein. The slope heuristics have several versions (see [Arl19]). In this paper we use the dimension jump algorithm, which is implemented for instance in the R package `capush`.

The aim is to tune the constant κ in a penalty of the form $\operatorname{pen}(m) = \kappa \operatorname{pen}_{\text{shape}}(m)$ where $\operatorname{pen}_{\text{shape}}$ is a known penalty shape. In the most standard cases, $\operatorname{pen}_{\text{shape}}$ is the dimension of the model. Let $\hat{m}(\kappa)$ be the model selected by the penalized criterion with constant κ

$$\hat{m}(\kappa) \in \operatorname{argmin}_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|Y - \hat{f}_m\right\|_n^2 + \kappa \operatorname{pen}_{\text{shape}}(m)\right\}.$$

The Dimension Jump algorithm consists of the following steps (see Figure 3b for an illustration)

1. Compute $\kappa \mapsto \hat{m}(\kappa)$,
2. Find the constant $\hat{\kappa}^{dj} > 0$ that corresponds to the highest jump of the function $\kappa \to d_{\hat{m}(\kappa)}$,
3. Select the model $\hat{m}(2\hat{\kappa}^{dj})$,

$$\hat{m} \in \operatorname{argmin}_{m\in\mathcal{M}}\left\{\left\|Y - \hat{f}_m\right\|_n^2 + 2\hat{\kappa}^{dj} \operatorname{pen}_{\text{shape}}(m)\right\}.$$

4.2 Presentation of the experiments

We simulate n observations according to the following generative model on $[0,1]$

$$Y_i = f^*\left(\frac{i}{n}\right) + \varepsilon_i, \quad i = 1 \dots n. \tag{4.1}$$

In the simulations we take for f^* the function

$$f^* : t \in [0,1] \mapsto 3 - 0.1 * t + 0.5 * t^2 - t^3 + \sin(8 * t).$$

The aim is to estimate f^* on a regular partition of size m, for $m \in \{1, \dots, 200\}$. We simulate n observations ε according to an ARMA process, a Fractional Gaussian process and a non Gaussian Markov chain. The last framework allows us to evaluate the robustness of the model selection procedure without the Gaussian assumption. We shall consider samples of size $n = 200$, $n = 500$, $n = 2000$, $n = 5000$; for each estimator that will be considered below, the boxplots of the risks will be carried out through 100 independent trials.

We now give more details on the error processes we use for the simulations.

- **ARMA process.** The ARMA(2,1) short memory process is defined by

$$\varepsilon_i - 0.3\varepsilon_{i-1} - 0.1\varepsilon_{i-2} = W_i + 0.2W_{i-1}, \tag{4.2}$$

  where $(W_i)_{i\in\mathbb{Z}}$ is a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables.

- **Fractional Gaussian Noise.** The Fractional Gaussian Noise (FGN, see for instance [MVN68] and [Ber94]) is a stationary sequence $(\varepsilon_i)_{i\geq 1}$ of zero-mean Gaussian random variables with auto–covariances

$$\gamma_\varepsilon(k) = \frac{\sigma^2}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right), \quad \text{for } k \in \mathbb{N},$$

  where $\sigma^2 = \gamma_\varepsilon(0) = \mathrm{Var}(\varepsilon_i)$, and $H \in (0,1)$ is the so-called Hurst parameter. If $H = 1/2$, the sequence $(\varepsilon_i)_{i\geq 1}$ is a Gaussian white noise with variance σ^2. For any $H \in (0,1)$ the following asymptotic expansion is valid

$$\gamma_\varepsilon(k) \sim \sigma^2 H(2H-1)k^{2(H-1)}\,.$$

  Consequently, if $H > 1/2$, the process is positively correlated and long-range dependent. If $H < 1/2$, the process is negatively correlated and $\sum_{k\geq 0}|\gamma_\varepsilon(k)| < \infty$, so that (3.1) holds and the process is short-range dependent.

  In fact, for $H < 1/2$, the FGN $(\varepsilon_i)_{i\geq 1}$ is anti-persistent in the sense of Definition 3.5 (with $\gamma = 2 - 2H$ in Definition 3.5). This is well known (see for instance [Ber94]), and follows from the fact that the ε_i's are the increments of a fractional Brownian motion B_H, that is for $i = 1, 2, \dots$

$$\varepsilon_i = B_H(i) - B_H(i-1), \quad \text{with } \mathrm{Var}(B_H(t)) = \sigma^2 t^{2H}.$$

  In the simulations, we shall consider two cases

  - an anti-persistent case, with $H = 0.2$,
  - a long memory case, with $H = 0.7$.

- **Non Gaussian Markov chain.** We start from the Markov chain introduced by Doukhan, Massart and Rio [DMR94].

  Let a be a positive real number, let ν be the probability with density $x \to (1+a)x^a \mathbf{1}_{[0,1]}$ and π be the probability with density $x \to ax^{a-1}\mathbf{1}_{[0,1]}$. We define now a strictly stationary Markov chain by specifying its transition probabilities $K(x, A)$ as follows

$$K(x, A) = (1-x)\delta_x(A) + x\nu(A)\,,$$

  where δ_x denotes the Dirac measure at point x. Then π is the unique invariant probability measure of the chain with transition probabilities $K(x,\cdot)$. Let $(Z_i)_{i\in\mathbb{Z}}$ be the stationary Markov chain on $[0,1]$ with transition probabilities $K(x,\cdot)$ and invariant distribution π. From [DMR94], we know that the

β-mixing coefficients $\beta_Z(n)$ of the chain are such that $\beta_Z(n) \sim \frac{1}{n^a}$. One can easily check than Z_i^a is uniformly distributed over $[0,1]$, so that

$$\varepsilon_i = Z_i^a - 0.5$$

is a stationary Markov chain (as an invertible function of a stationary Markov chain), with mean zero and mixing coefficient $\beta(k) \sim \frac{1}{n^a}$. This chain is short range dependent if $a > 1$ and long-range dependent if $a \in (0,1)$ (see for instance [DGM18] for a deeper discussion on this subject).

In the simulations, we shall consider three cases

- two short memory cases, with $a = 8$ and $a = 1.5$,
- a long memory case, with $a = 0.5$.

In fact, for regressograms on a regular partition of size m, the main term of the penalty can be exactly determined by the behavior of $\mathrm{Var}(\varepsilon_1 + \cdots + \varepsilon_n)$ (see the proof of Lemma 3.2). More precisely, if

$$\mathrm{Var}\left(\sum_{k=1}^{n} \varepsilon_k\right) \sim \kappa n^{2-\gamma}\,,$$

for some $\gamma \in (0,2)$, then the main term of the penalty will be of order $(m/n)^\gamma$. We then see that γ is related to the usual Hurst index H (see for instance [Ber94]) of the partial sum process

$$S_n = \varepsilon_1 + \cdots + \varepsilon_n\,,$$

via the equality $\gamma = 2 - 2H$. Hence, for regressograms on a regular partition of size m, the main term of the penalty is of order $(m/n)^{2-2H}$.

For long range dependent Gaussian processes, the variance terms of the risk are not linear functions of the dimension, for sufficiently large dimension they behave as m^γ for some $\gamma \in (0,1)$. Figure 1 shows the risk curve $m \to \mathbb{E}[\|f^* - \hat{f}_m\|_n^2]$ of the regressograms for observations simulated according to (4.1) with the error process following a Fractional Gaussian distribution with Hurst exponents between 0.1 and 0.9. As usual this risk is estimated via a basic Monte-Carlo procedure, by averaging the empirical risk $\|f^* - \hat{f}_m\|_n^2$ over $N = 100$ independent trials. For Figure 1, the size of the samples is equal to $n = 2000$, and m varies from 1 to 700.

For anti-persistent cases ($H < 0.5$), the risk curve has a convex behavior for large dimensions, in accordance with a variance term of order m^{2-2H} (see Section 3.3). For the i.i.d. case ($H = 0.5$), the risk curve is linear for high dimensions. For the long range dependent cases ($H > 0.5$), the risk curve shows a concave behavior for large dimensions, in accordance with a variance term of order m^{2-2H} (see Section 3.2).

Figure 2 shows the risk curve of the regressograms for observations simulated according to (4.1), when the error process is the β-mixing Markov chain described above with a parameter a between 0.3 and 10. We remark a concave behavior for large dimensions in the long range dependent case ($a < 1$) and a linear behavior for large dimensions in the short range dependent case ($a > 1$). This suggests that the theoretical results obtained in Sections 3.1 and 3.2 could be also valid in non Gaussian contexts.

For the simulations, we use the Whittle MLE-estimator [Whi53] implemented in the `longmemo` package, to estimate the Hurst index H. We compare several approaches

- **CDJ**: Classical Dimension Jump method with a penalty shape proportional to the dimension.
- **HGiven**: Dimension Jump for the penalty shape m^{2-2H} with Hurst exponent H given.
- **Wh(Y)**: Dimension Jump for the penalty shape $m^{2-2\hat{H}}$ where $\hat{H}$ is the Whittle estimator computed on the Y process.
- **Wh(Res)**: Dimension Jump for the penalty shape $m^{2-2\hat{H}}$ where $\hat{H}$ is the Whittle estimator computed on the residuals of a model.
  For the method Wh(Res), we have to propose a model m_0 for which the Hurst exponent is computed on the residuals. Roughly speaking, the idea is to estimate the Hurst exponent in a sufficiently large model for which the bias is negligible. We propose a two steps procedure, which is based on the selection of a pre-model $\hat{m}_1$ to estimate the Hurst exponent H on the residuals of $\hat{m}_1$. This provides an estimator $\hat{H}$ which is used to design the penalty shape. The dimension jump is then used to select the final model $\hat{m}$. We propose two versions for this two-step procedure:

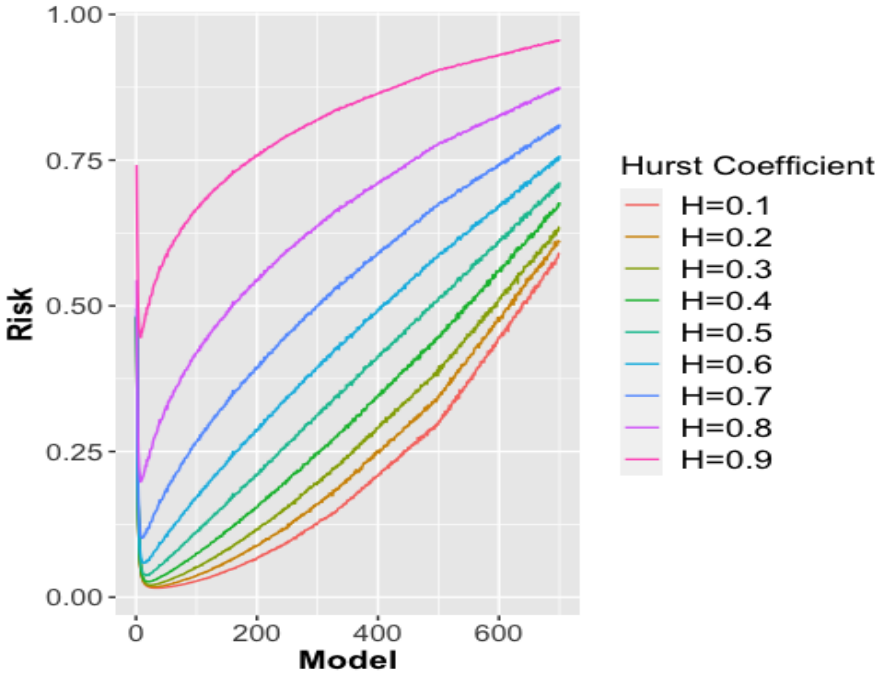


Figure 1: Comparison of risk shapes for the fractional Gaussian process with Hurst coefficient between 0.1 and 0.9, and for $n = 2000$.

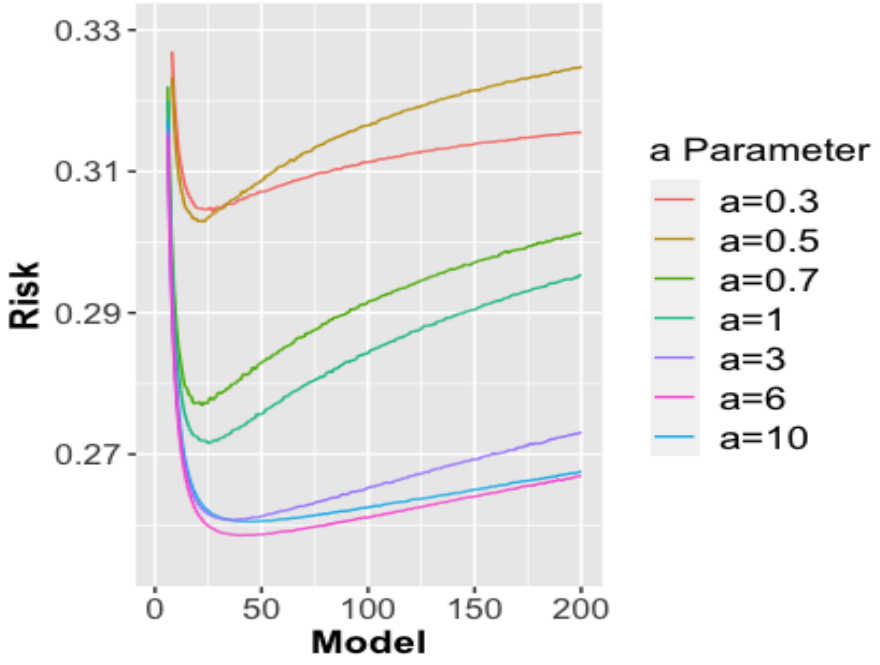
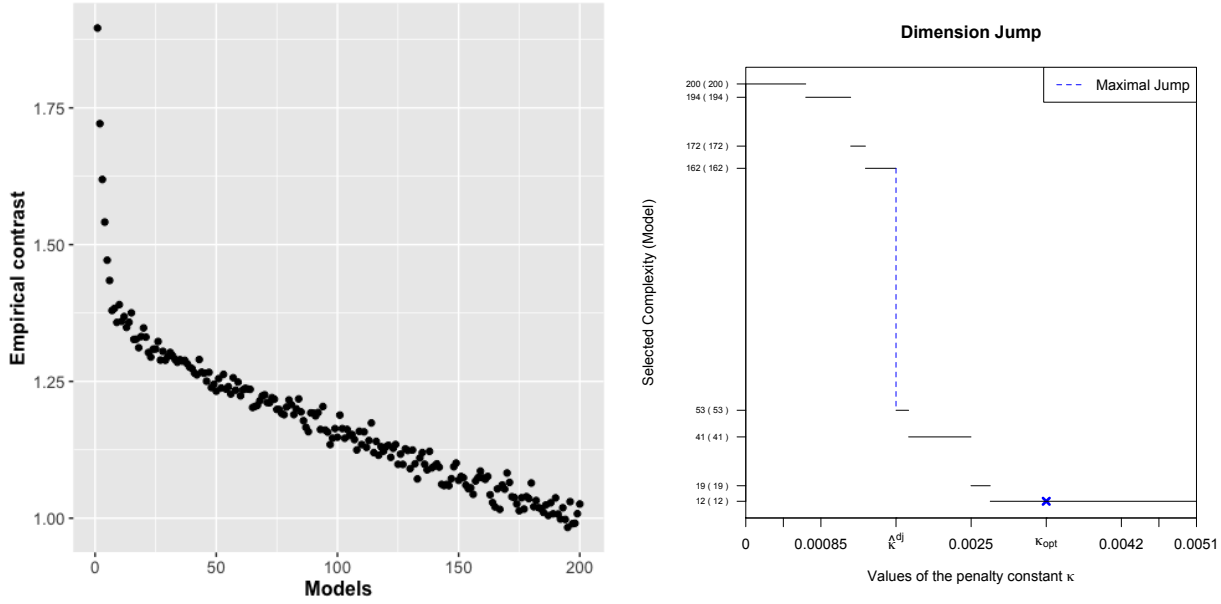


Figure 2: Comparison of risk shapes for the Markov chain, for $n = 2000$.



(a) Linear behavior of the empirical contrast ($n = 2000$).

(b) Dimension Jump.

Figure 3: Illustration of the slope heuristics for the ARMA(2,1) process.

- **CDJ+Wh(Res)**: Classical Dimension Jump to find a pre-model $\hat{m}_1$, then Whittle estimator $\hat{H}$ to estimate the Hurst exponent and finally Dimension Jump with penalty shape $m^{2-2\hat{H}}$.
- **Wh(Y)+Wh(Res)**: Dimension Jump with penalty shape $m^{2-2\hat{H}_1}$ where $\hat{H}_1$ is the Whittle estimator on Y, this selects a pre-model $\hat{m}_1$, then Whittle estimator $\hat{H}_2$ on the residuals of the model $\hat{m}_1$ and finally Dimension Jump with penalty shape $m^{2-2\hat{H}_2}$.

4.3 Short range dependence

In this section we study the performance of the model selection method in the short dependence framework. The penalty shape is chosen proportional to the model dimension, as in the i.i.d. case and we can apply the classical dimension jump method (CDJ) to calibrate κ. Roughly speaking, the slope heuristics relies, among other assumptions, on the fact that the empirical contrast behaves in high dimension as a linear function of the penalty shape.

We also compare the performances of the CDJ method with the ones of the other approaches. As we shall see, other methods can give better results for n small.

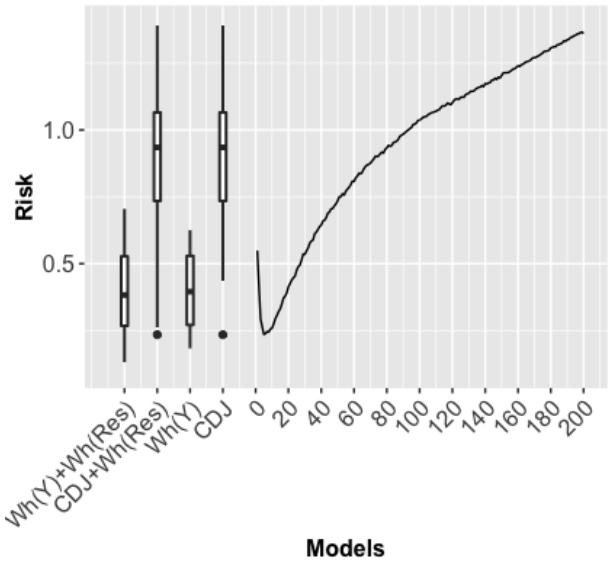
• Gaussian ARMA process

We begin with the classical ARMA(2,1) short memory process defined in (4.2). Figure 3 shows the behavior of the empirical contrast for $n = 2000$ and an illustration of the dimension jump algorithm. As expected by the slope heuristics, a linear behavior of the empirical contrast can be observed in high dimensions ($m \geq 25$).

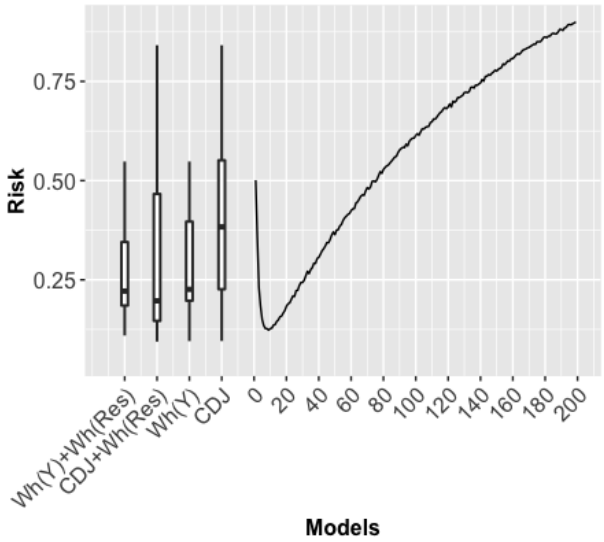
Figure 4 shows the performance of the different methods. The boxplots on the left part of each graph show the risk of this model selection method over 100 trials. On the right, the risk function is displayed.

In this experiment, the classical dimension jump (penalty shape proportional to the dimension) works clearly well for n large ($n \geq 2000$). It is however less efficient for n small. Indeed, the risk shows a concave behavior in large dimensions, as in the long memory case (as we shall see later on). For small n, an estimation of H with the Whittle estimator applied on the Y process and plugged into the penalty shapes finally gives better results than the classical dimension jump method.

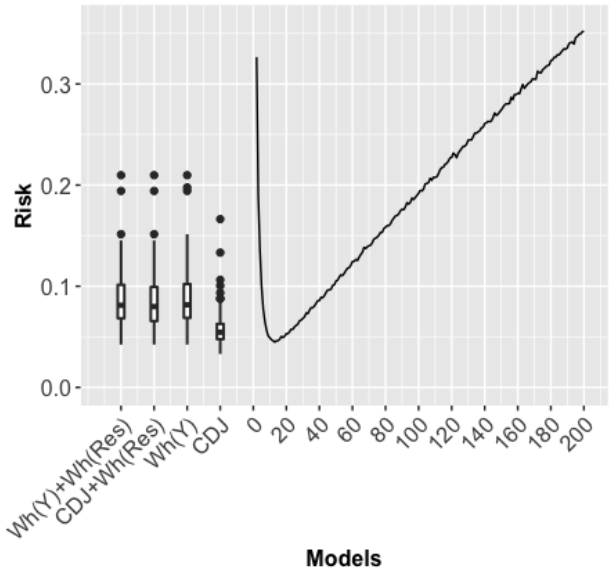
The Whittle estimator computed on the residuals is also efficient for selecting the minimal risk model for n small. In this case we consider the residuals process of the model chosen at first step either by CDJ or by Wh(Y), the method CDJ + Wh(res) having bad results for n too small ($n = 200$).



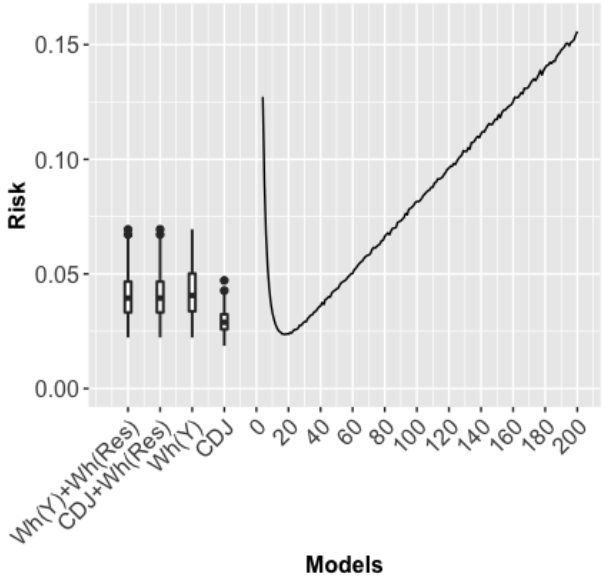
(a) $n = 200$



(b) $n = 500$

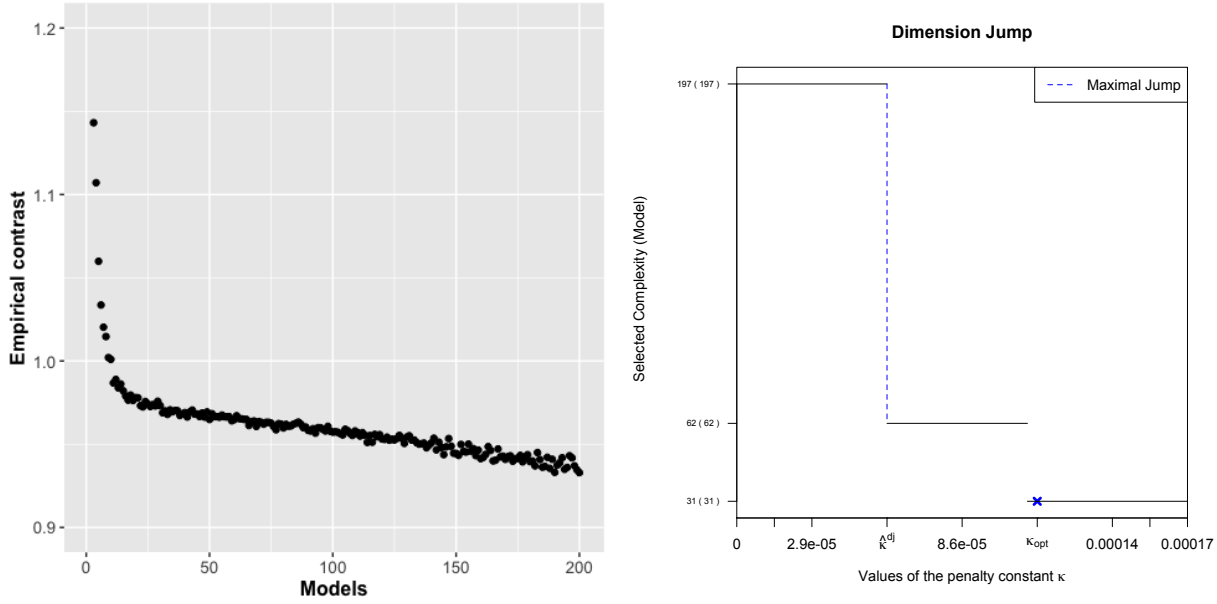


(c) $n = 2000$



(d) $n = 5000$

Figure 4: Short Memory ARMA process. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.



(a) Linear behavior of the empirical contrast.

(b) Dimension Jump.

Figure 5: Illustration of the slope heuristics for the non Gaussian process ($a = 8$).

• Non Gaussian Markov chain

To evaluate the robustness of the model selection procedure without the Gaussian error assumption, we consider the Non Gaussian Markov chain defined above. We simulate an error process ε distributed according to this stationary Markov chain, and we first make simulations in the short dependent case with a value of $a = 8$. As shown by Figure 5, a linear behavior of the empirical contrast can be observed, which is a good point for applying the slope heuristics here.

The performances of the methods are summarized on Figure 6. We can check on this figure that the classical dimension jump shows good performances. For all sample sizes, the dimension jump based on the Whittle estimator applied to Y is a little less efficient than the two-step methods.

We now consider a second short memory case with the Markov chain, with $a = 1.5$. This case is very close to the limit case $a = 1$, which separates long memory from short memory. Figure 7 shows that the CDJ method works well for n large. But for n small, the four methods do not really manage to select a model close to the oracle model.

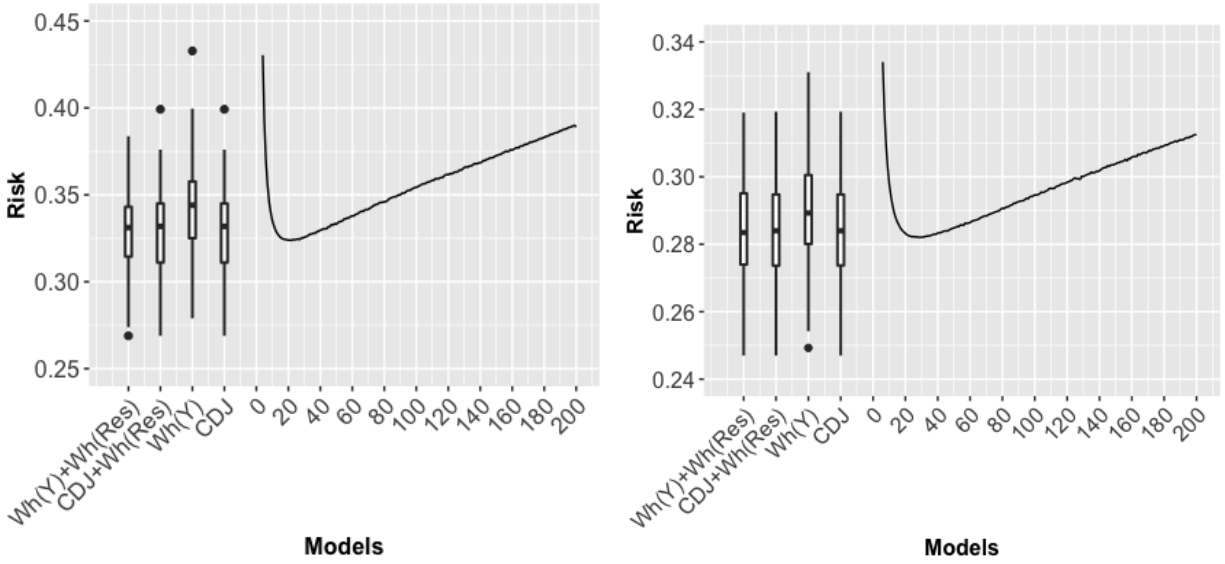
The methods based on the direct estimation of the Hurst exponent, like Wh(Y), give good results for n smaller than 500. Regarding the two-step methods, CDJ+Wh(res) shows bad performances for n small ($n \leq 500$), while Wh(Y)+Wh(res) shows good results for $n = 500$ but poor results for $n = 200$.

4.4 Long range dependence

For long range dependent Gaussian processes, the variance terms of the risk are not linear functions of the dimension, they behave as m^γ for some parameter $\gamma \in (0, 1)$. We thus would like to use penalties proportional to m^γ, see Section 3.2. For instance, for Fractional Gaussian processes, $\gamma = 2 - 2H$, where H is the Hurst exponent. Of course this coefficient is unknown in practice and thus we use some estimator of the Hurst exponent to calibrate the penalty. Generally speaking, estimating the Hurst exponent is a difficult statistical task, however a rough estimation can be sufficient for the model selection problem we study here.

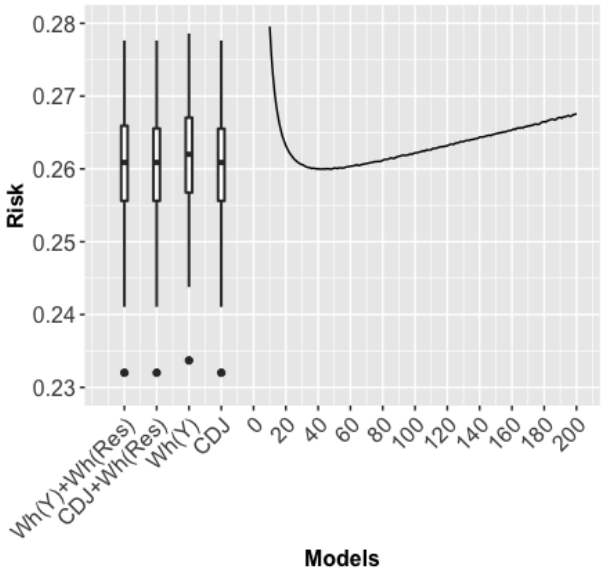
• Fractional Gaussian Noise

For this experiment we simulate the error process with a Gaussian Fractional Noise of Hurst parameter $H = 0.7$. The performances of the methods are summarized on Figure 8. We can check on this figure that when using a penalty with the true Hurst exponent ($H = 0.7$) of the error process, the model selection method works correctly. We also note that the classical dimension jump (penalty shape proportional to the dimension) shows bad performances. On the other hand, the Whittle estimators applied to Y and plugged into the penalty shape show good results for all sample sizes. The two steps methods show also good per-

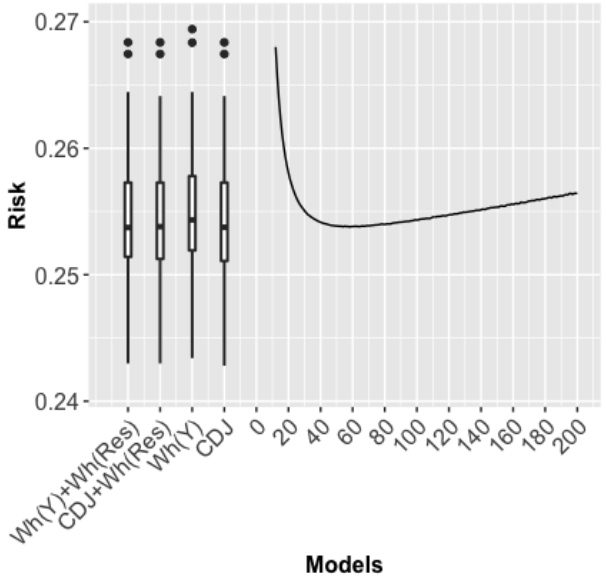


(a) $n = 200$

(b) $n = 500$

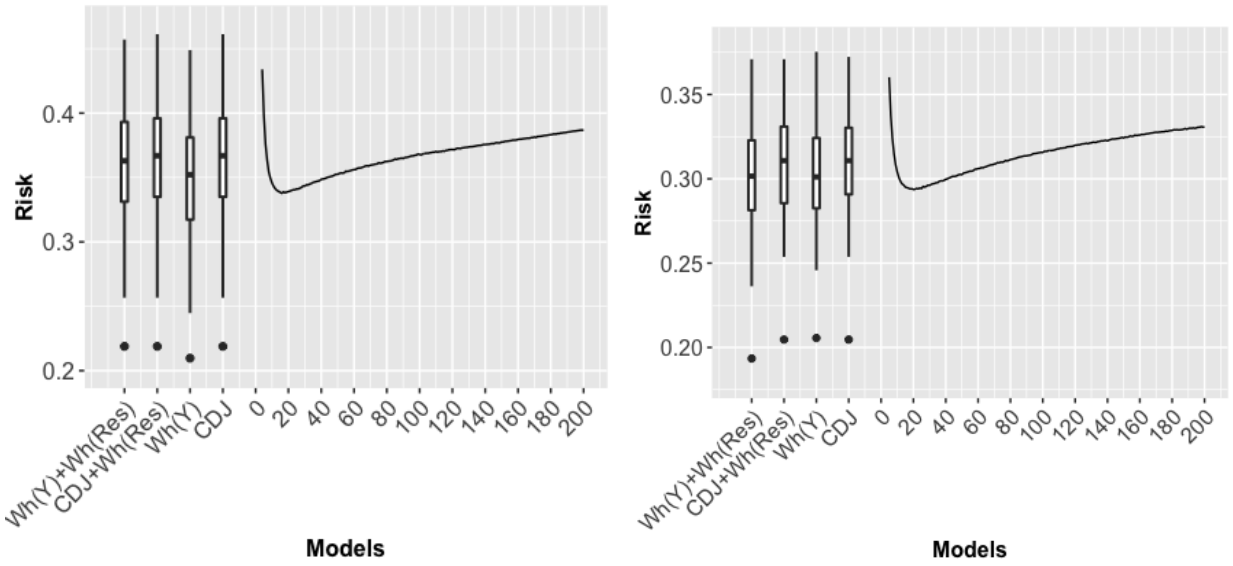


(c) $n = 2000$



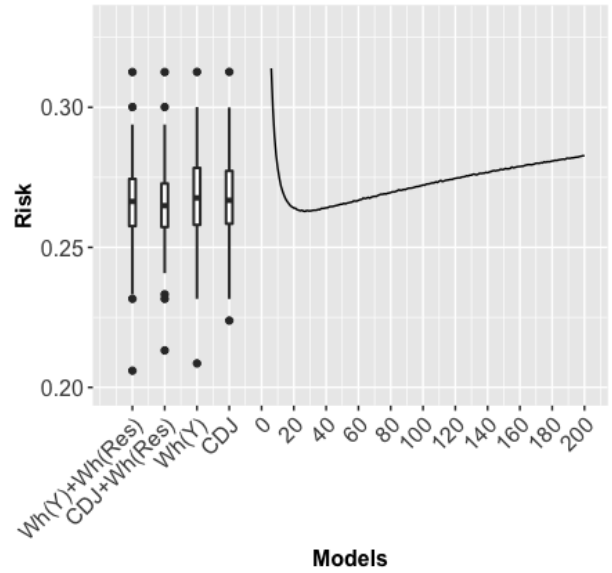
(d) $n = 5000$

Figure 6: Markov chain with $a = 8$. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.

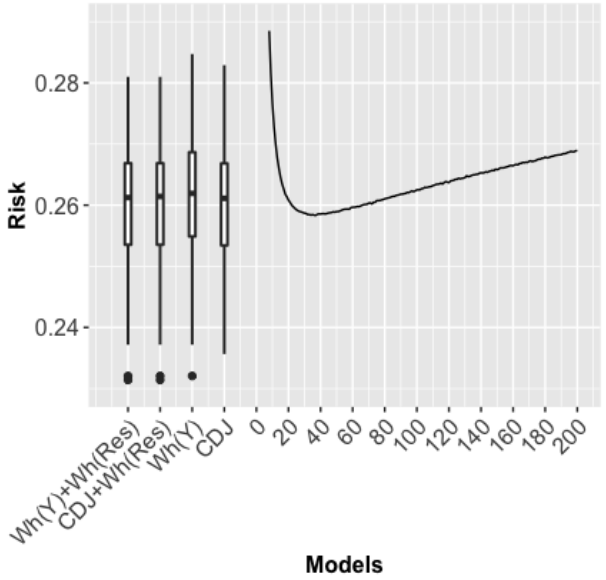


(a) $n = 200$

(b) $n = 500$



(c) $n = 2000$



(d) $n = 5000$

Figure 7: Markov chain with $a = 1.5$. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.

formances for n large enough.

- **Non Gaussian Markov chain**

We now evaluate the robustness of our model selection procedure when the Gaussian error assumption is not satisfied. We consider here the Non Gaussian Markov chain in the long range dependent setting. As for the Fractional Gaussian Noise, the risk has a concave behavior for large dimension, see Figure 2 for an illustration. Then the penalty shape is equal to m^a, where a is the decay rate of the covariances.

For this experiment we simulate the Markov chain with $a = 0.5$ for the error process. The performances of the methods are displayed on Figure 9. We observe that the classical dimension jump shows bad performances in this non Gaussian long range dependent context. When using the penalty shape m^a (H given, with $a = 2 - 2H$), the performances are a little better than before, but not as good as one could hoped for. For n large enough ($n \geq 2000$), the Whittle estimator applied on Y and plugged into the penalty shape shows satisfactory results. The performances of the two steps methods are similar but from $n = 5000$ only.

This experiment suggests that more work should be done in this context. It seems that a concave penalty shape should be used, as expected, but that the good exponent could perhaps be different from $a = 2 - 2H$.

4.5 Anti-persistent errors with a Fractional Gaussian Noise

We consider the same simulation protocol with anti-persistent errors, following a Fractional Gaussian Noise with $H = 0.2$. Again, we observe a linear behaviour of the empirical contrast in high dimension, see Figure 11a.

The performances of the different methods on this experiment are summarized by Figure 10. We can check that when using a penalty with the true Hurst exponent ($H = 0.2$), the model selection method works pretty well. The two-step methods, with the Whittle estimator computed on the residuals, give similar results for all n. On the other hand, the Whittle estimator applied directly on Y shows poor performances for n small, but it is better for n large.

We also note that, in this short range dependent case, the classical dimension jump shows good results for all n, as in the i.i.d. case.

4.6 Impact of long memory on risk performances

The aim of this experiment is to discuss the impact of long memory on the risk performance of our model selection procedure. We consider $n = 2000$ observations according to the generative model defined by Equation (4.1) with a Fractional Gaussian noise distribution. We study the risk performances of three methods for H between 0.5 (i.i.d. case) and 0.9 with both the Wh(Y)+Wh(Res) method and CDJ method. The boxplots of the risks are carried out through 100 independent trials.

The results are summarized in Figure 12. Of course, for both methods the larger H, the more the risk performances deteriorate. We also see that, compared with the Wh(Y)+Wh(Res) method, the risk of the CDJ procedure strongly increases for long memory.

We have also estimated the regression function by a naive kernel method with a cross validation procedure to choose the bandwidth parameter. Since we have use regressograms in our model selection procedure, it is natural to make the comparison with the uniform kernel. We apply the `npreg` and `npregbw` functions of the R package `np` which implements a least square cross validation to choose the bandwidth parameter (default method). Figure 12 shows that the the naive uniform kernel method with cross validation performs badly for long memory, and this is in spite of the fact that it performs really better than the regressogram for independent observations ($H = 1/2$).

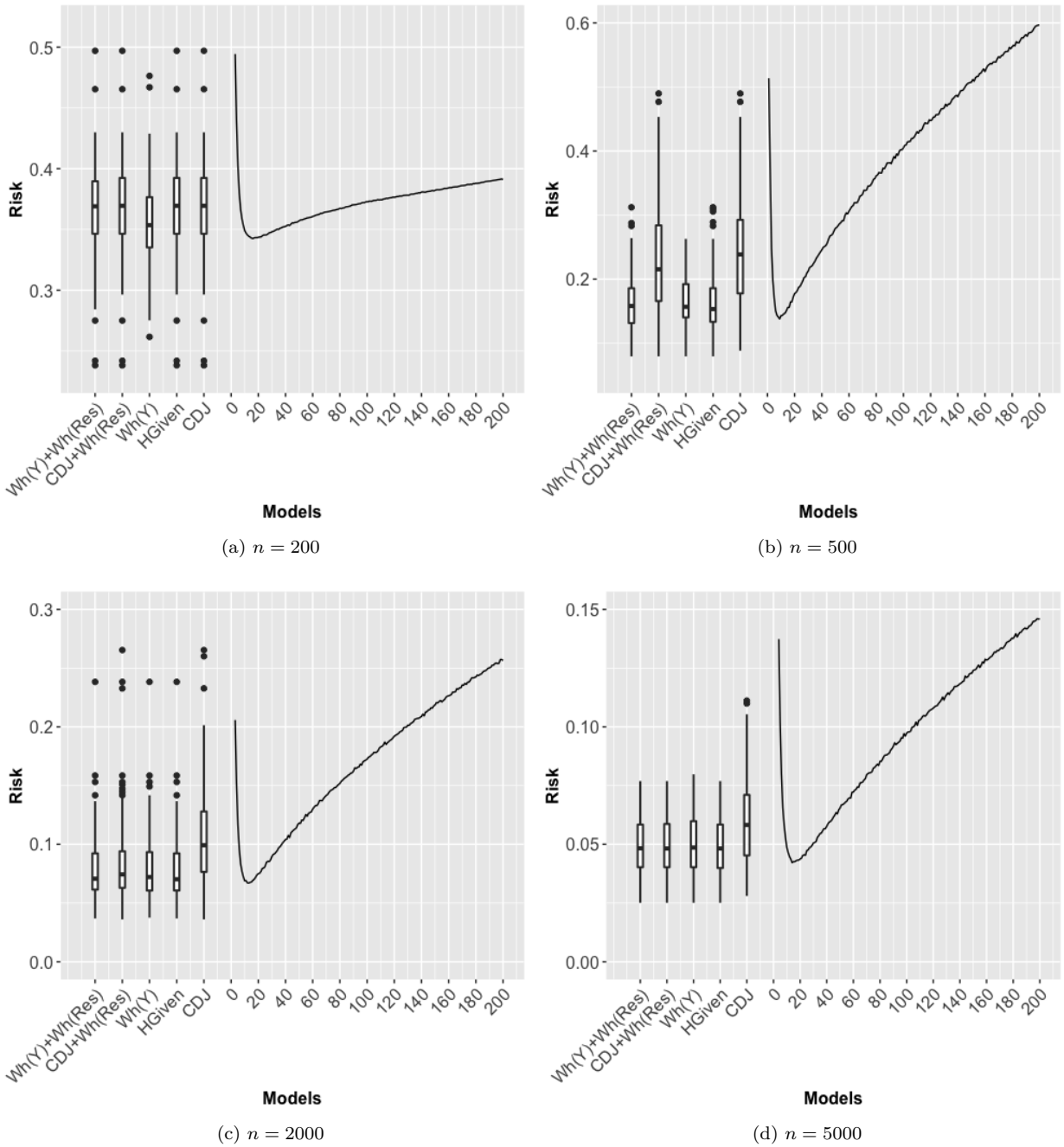
In conclusion, for long memory settings our method Wh(Y)+Wh(Res) should be preferably used rather than the CDJ method or a standard kernel method.

4.7 Identification

In this experiment, we illustrate some satisfactory properties of our model selection procedure in the specific case where the true regression function belongs to the model family. We consider the regression model given by Equation (4.1) with

$$f^\star : x \in [0,1] \mapsto \frac{1}{2} \sum_{j=0}^{5} (-1)^j \mathbb{1}_{[\frac{j-1}{6}, \frac{j}{6}[}(x).$$

For the error we simulate a Fractional Gaussian noise process with $H = 0.8$.



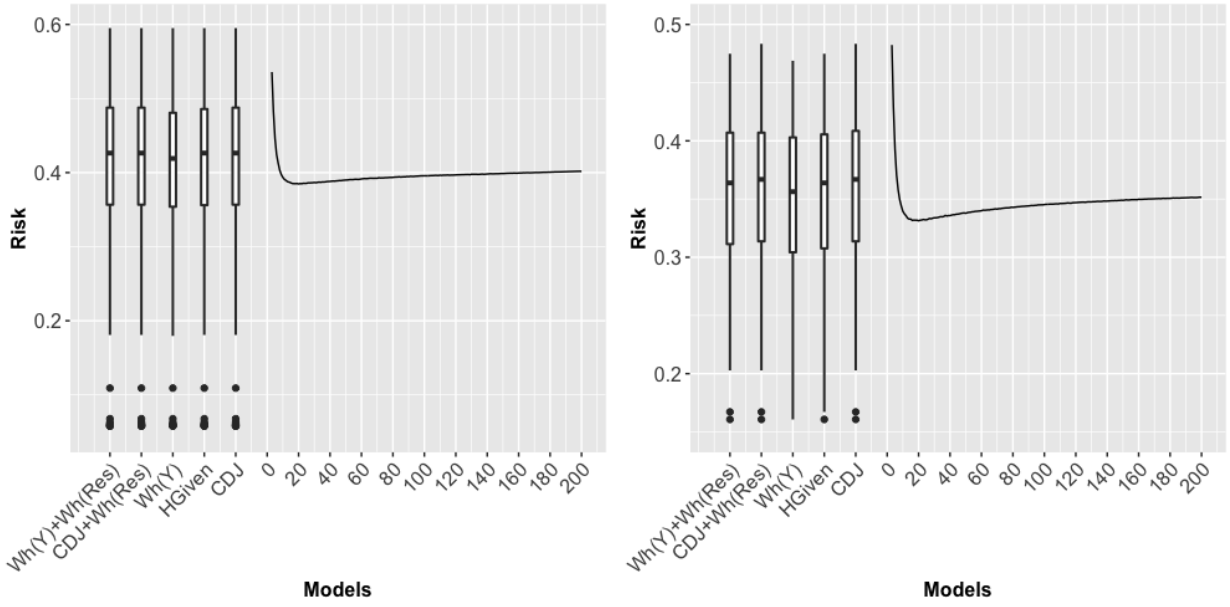
(a) $n = 200$

(b) $n = 500$

(c) $n = 2000$

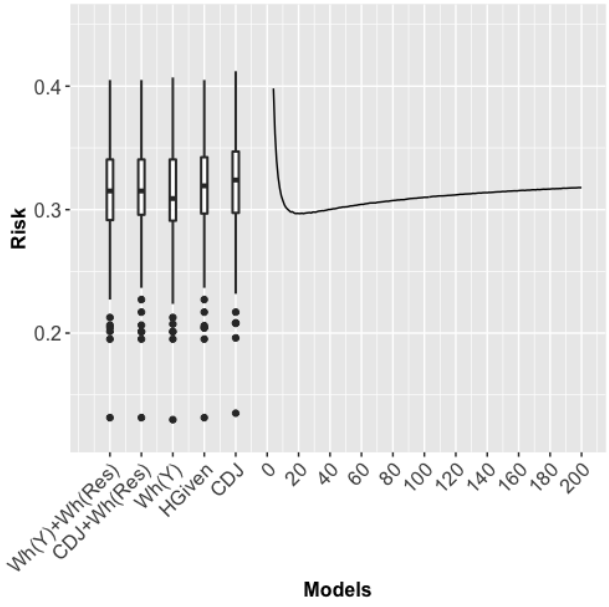
(d) $n = 5000$

Figure 8: Long Memory Fractional Gaussian error process with $H = 0.7$. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.

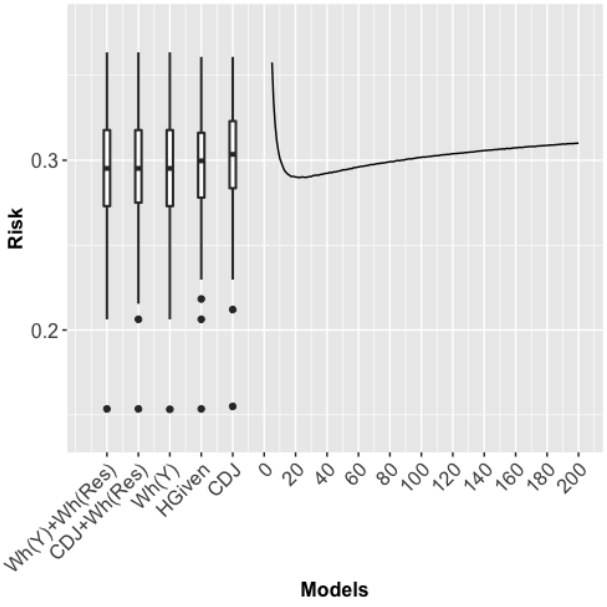


(a) $n = 200$

(b) $n = 500$

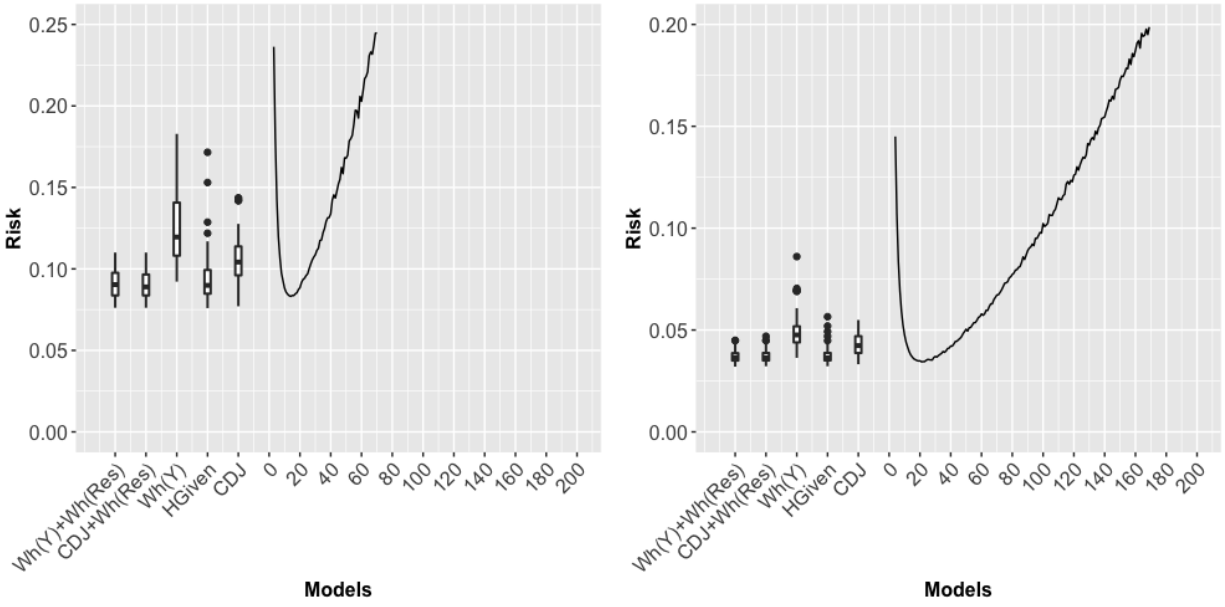


(c) $n = 2000$



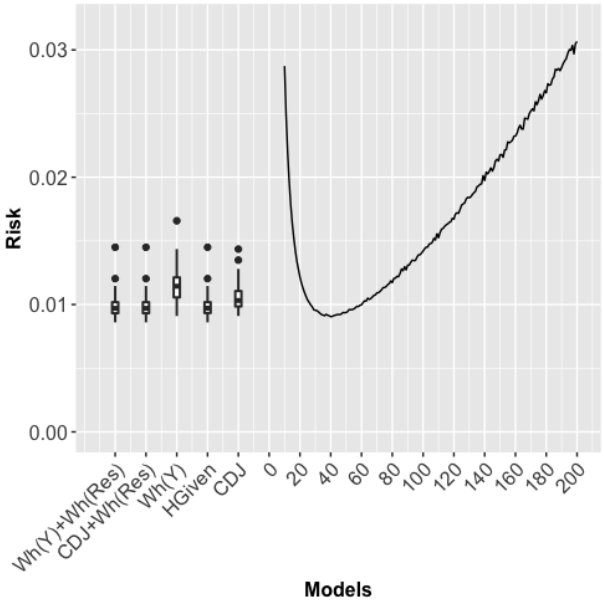
(d) $n = 5000$

Figure 9: Markov chain process with $a = 0.5$. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.

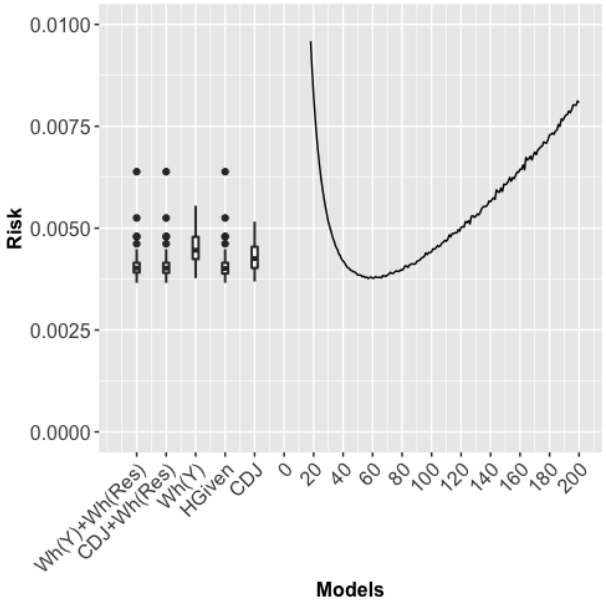


(a) $n = 200$

(b) $n = 500$

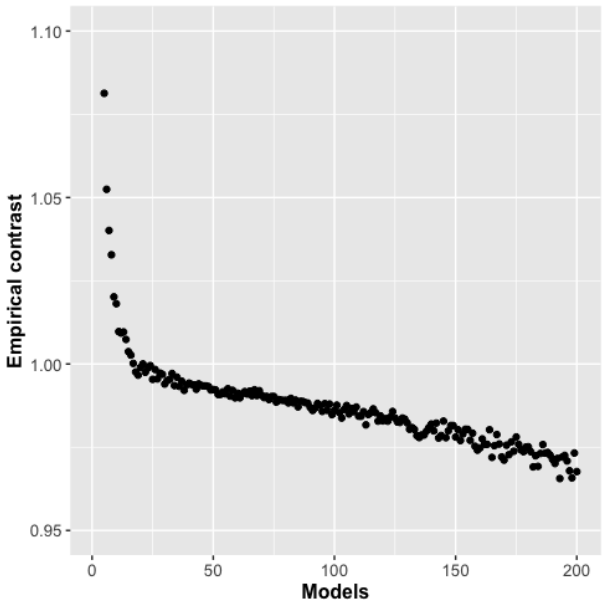


(c) $n = 2000$

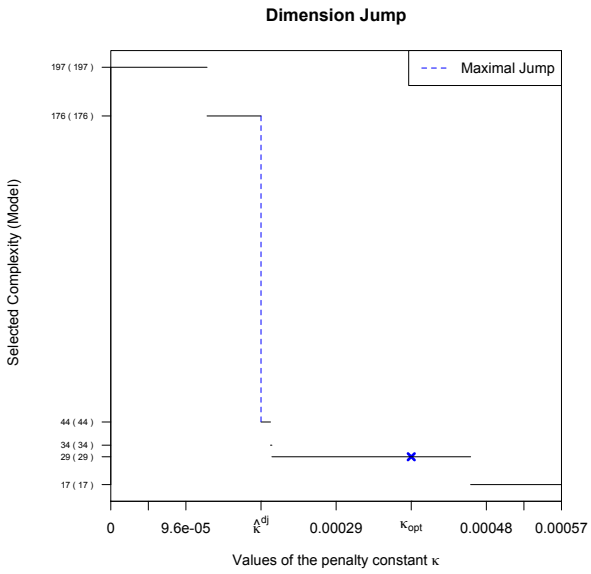


(d) $n = 5000$

Figure 10: Short Memory Fractional Gaussian process with $H = 0.2$. Risk curves and performances of the different calibration methods for $n = 200, 500, 2000, 5000$.



(a) Linear behavior of the empirical contrast for $n = 2000$.



(b) Dimension Jump.

Figure 11: Illustration of the slope heuristics for the Fractional Gaussian process ($H = 0.2$).

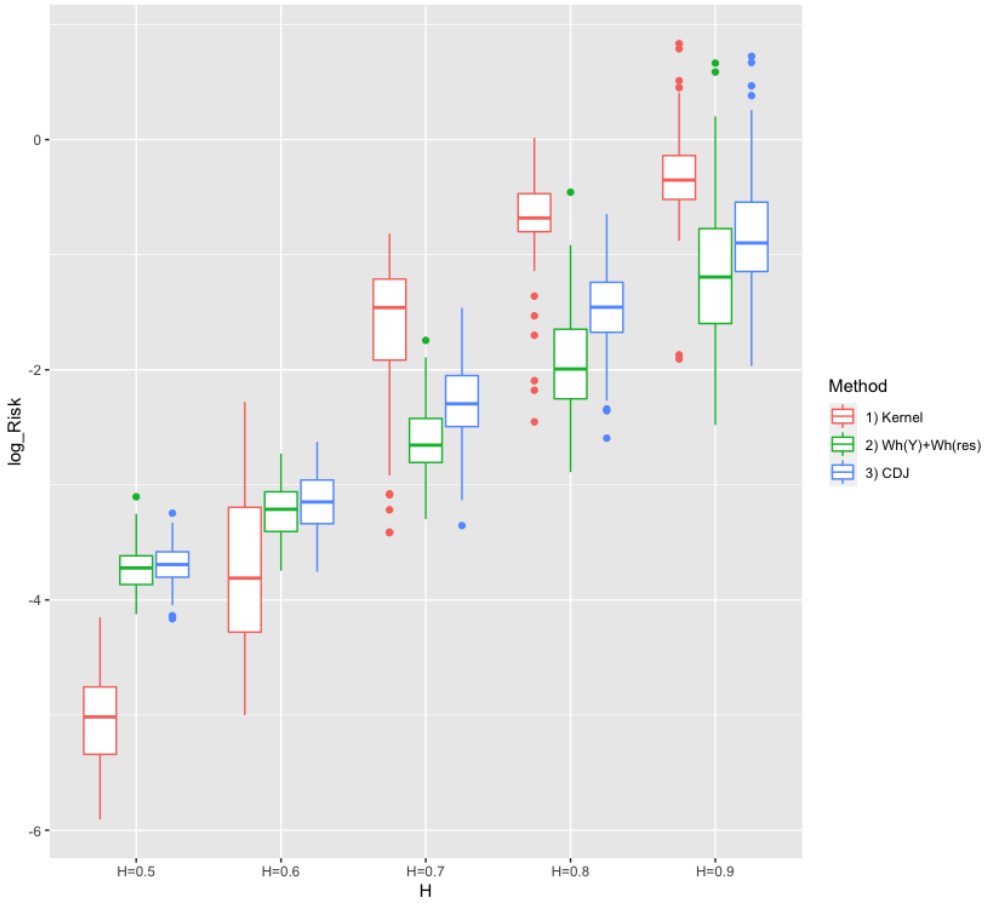


Figure 12: Boxplots of risk performances logarithmic scale for three methods with H between 0.5 and 0.9.

n	Method	< 3	4	5	**6**	7	8-11	12	> 12
500	Wh(Y)+Wh(Res)	16	0	0	**71**	0	3	8	2
	CDJ+Wh(Res)	0	0	0	**26**	0	1	11	62
	Wh(Y)	16	0	0	**82**	0	1	1	0
	CDJ	0	0	0	**4**	0	0	5	91
5000	Wh(Y)+Wh(Res)	0	0	0	**90**	0	0	10	0
	CDJ+Wh(Res)	0	0	0	**82**	0	0	14	4
	Wh(Y)	0	0	0	**93**	0	0	7	0
	CDJ	0	0	0	**2**	0	0	5	93

Table 1: Identification of the correct model in the collection by the model selection procedures over 100 simulations.

The results shown in this paper do not concern the identification of the correct model in the family. However Table 1 suggests that our model selection procedures easily identify the correct true model in the collection. This is not the case with the CDJ procedure, which very often selects a model that is far too large.

4.8 Conclusion on the experiments

In these experiments we see that the penalty proportional to (m/n) (with a constant calibrated thanks to the jump dimension algorithm: CDJ method) performs quite well for short memory processes, but underperforms in all the other situations. The Wh(Y) method, with a penalty proportional to $(m/n)^{2-2\hat{H}}$ and an estimator $\hat{H}$ based ont he Y_i's, performs quite well in most of the cases, but can show very bad performances (see for instance Figure 10) and is hard to justify from a heuristic point of view. The two steps methods, with a penalty proportional to $(m/n)^{2-2\hat{H}_2}$ and an estimator $\hat{H}_2$ based on the residuals of the first adjustment, perform well in most of the cases, with a clear preference for the Wh(Y)+Wh(Res) method. In fact, we suspect an overfitting with method CDJ for long memory processes, so that the residuals based on CDJ are not close to the original error process (see Table 1 and also the application to the Nile data in Section 5).

We note that the two steps method Wh(Y)+Wh(Res) gives performances close, even sometimes better, to the best of the other proposed methods. An interesting example is the Gaussian ARMA process: for large n ($n \geq 2000$), the risk curve is quasi linear, and the CDJ method is the best method. But for small n ($n \leq 500$), the risk curve is concave, as in the long memory case, and the Wh(Y)+Wh(Res) is the best method. This suggests that, even for short memory processes, a penalty proportional to (m/n) is not always a wise choice in practice.

Our final comment is then: instead of looking for a penalty proportional to $(m/n)^\gamma$ for an appropriate γ, it might be preferable to estimate directly the term $\mathrm{tr}(\mathrm{Proj}_{S_m}\Sigma)$. This could perhaps be done by giving an estimation of the covariance Σ based on the residuals of an appropriate pre-model.

5 Application to Nile data

In this section, we wish to continue the discussion on the Nile data initiated by Robinson in his 1997 article [Rob97]. We borrow from Robinson his presentation of this dataset, as well as some other sentences: "These data consist of readings of annual minimum levels at the Roda gorge near Cairo, commencing in the year 622; often only the first 663 observations are employed because missing observations occur after the year 1284 (see [Tou25]). It was one of the hydrological series examined by Hurst [Hur51] which led to his recognition of the "Hurst effect" and invention of the R/S statistic". The data are plotted in Figure 13.

Robinson then summarizes the different ways of understanding these data: either by considering that the cyclical variations come from a phenomenon of long memory, or by considering that the series can be written as the sum of a deterministic tendency plus a random noise. We refer to his article for relevant references on these questions.

Robinson applied different kernel estimators (with different bandwidths) to estimate the regression function. Then he estimated the Hurst coefficient H of the errors from the residuals of the regression (see Section 4 of his paper for the definition of the estimator of H). He noted that "These estimates thus vary greatly over the ranges of the smoothing employed" and concluded this section by "This study highlights the need for developing methods for choosing b and c which respond automatically to the strength of the dependence in u_t" (here b and c are the bandwidth used to estimate the regression function and the Hurst index respectively; u_t is the error process, according to Robinson's notations).

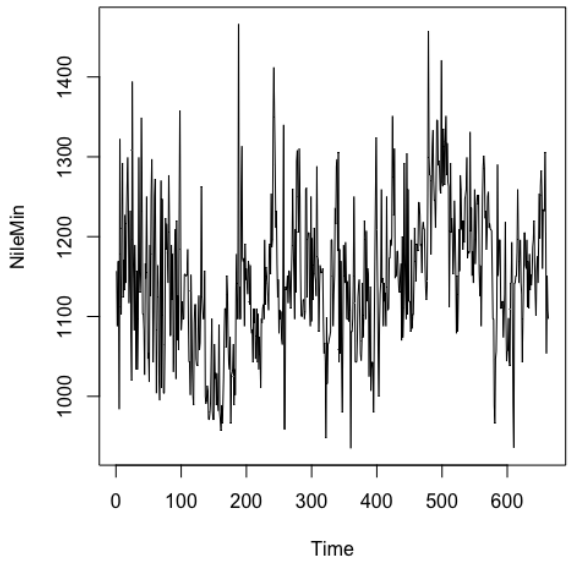


Figure 13: Nile River data.

This last sentence motivates us to apply our methods on these data, since we have a way to select automatically a partition from the data. We try two penalties: the usual penalty proportional to m/n, using the "classical jump dimension" to calibrate the constant (see CDJ method in Section 4); this method should work well if the underlying error process was short range dependent. And a penalty proportional to $(m/n)^{2-2\hat{H}_2}$, where $\hat{H}_2$ is the Hurst estimator based on the residuals, according to the Wh(Y)+Wh(Res) method described in Section 4. Indeed, this method was the best method according to the different kind of simulations done in Section 4. The resulting estimators are plotted in Figure 14.

The CDJ method selects a partition of size $m = 54$, with a clear impression of overfitting: the estimated trend seems very irregular, with many sudden changes. It seems that some randomness is still present in the trend. The Hurst index estimated through the residuals obtained with the estimated trend gives $\hat{H} = 0.59$, hence not so far from a white noise.

The Wh(Y)+Wh(Res) selects a much smaller partition, with $m = 7$. The trend looks more regular and interpretable, with a clear minimal period, a clear maximal period, and an almost constant tendency in between. It also suggests that an irregular partition should be used, which is a priori doable with our model-selection method, at the price of more tricky computations and algorithms. The Hurst index estimated through the residuals obtained with the estimated trend gives $\hat{H} = 0.79$, in accordance with the long-range dependence hypothesis.

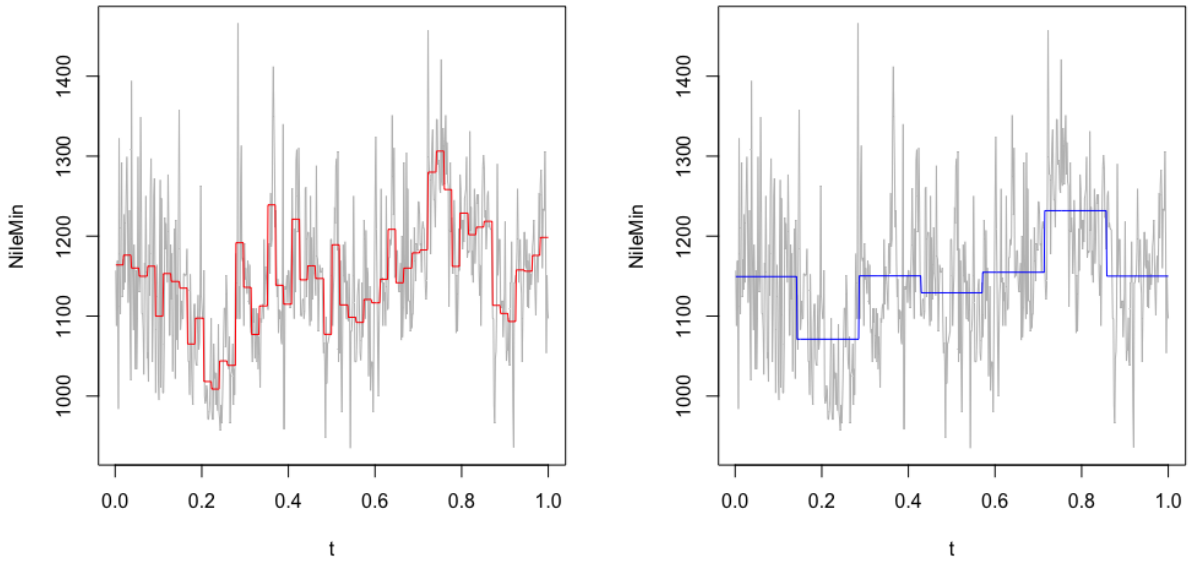
To be complete, the graph and the ACF of the residuals obtained with the Wh(Y) + Wh(Res) method are plotted in Figure 15.

6 Discussion

This paper deals with linear model selection with Gaussian dependent errors through ℓ_0 penalization. Several generalizations and extensions could be proposed in future works.

We apply Theorem 2.1 to study the fixed design case, but clearly the theorem also applies to all the settings considered in [BM01a] (or Chapter 2 in [Gir14]) in the i.i.d case. In particular, if the error process is short range dependent, then for all these problems the penalty is the same as in the i.i.d. case, the usual variance term being replaced by the spectral radius of the covariance matrix. One natural application concerns high dimensional problems and minimax rates of convergence in this setting, by considering for instance the models presented in Chapter 2 of [Gir14].

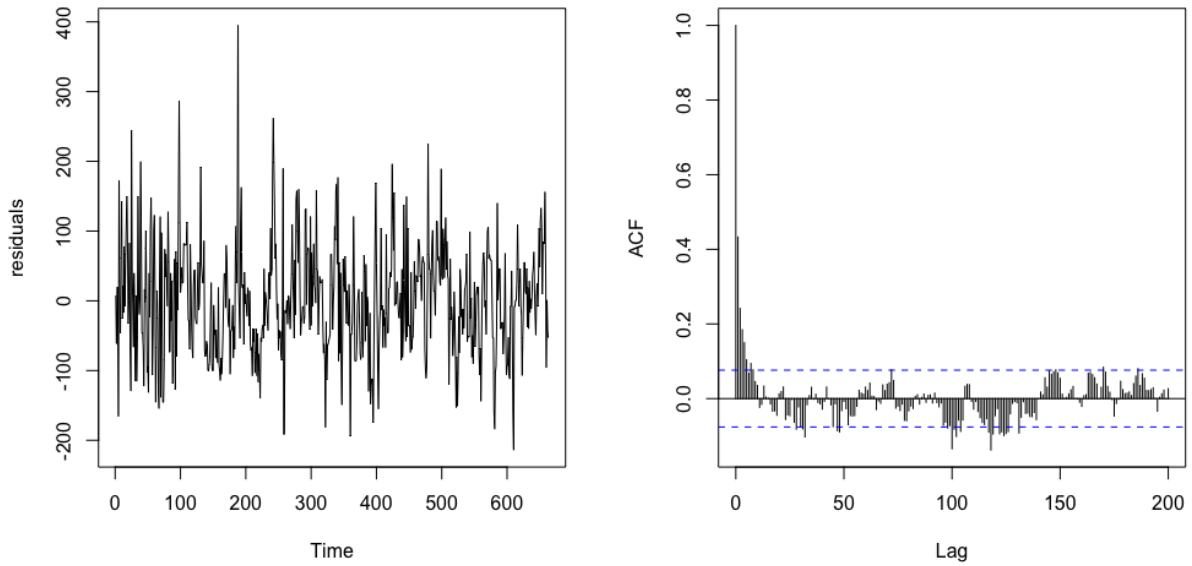
For long range dependent processes, the situation is more complicated, and a penalty involving the spectral radius only is no longer adapted. In that case, however, we conjecture that the upper bound given



(a) Regressogram with CDJ

(b) Regressogram with Wh(Y)+Wh(res)

Figure 14: Nile River data and resulting estimators.



(a) Residuals

(b) ACF

Figure 15: Residuals and ACF of the residuals for the method Wh(Y)+Wh(res).

in (2.4)

$$\operatorname{tr}\left(\operatorname{Proj}_{S_m}\Sigma\right) \leq \sum_{i=1}^{d_m} \lambda_i$$

should be of the right order. More precisely we formulate the following:

Conjecture. If $|\gamma_\varepsilon(k)| \leq \kappa k^{-\gamma}$ for some $\kappa > 0$ and $\gamma \in (0,1)$, then there exists a positive constant C such that

$$\sum_{i=1}^{d_m} \lambda_i \leq C d_m^{\gamma} n^{1-\gamma}\,. \tag{6.1}$$

This is an important conjecture: if (6.1) is true, it gives a practical way to deal with all the problems described above for long range dependent sequences, in the same way as we did with the fixed design regression model.

The performances of the ℓ_0 penalization strategy are studied in this work assuming that the distribution of the errors is stationary. However, Theorem 2.1 does not require this assumption. In a similar line of work, [Gen08] considers model selection for heteroscedastic Gaussian regression, for independent observations. It would be possible to study model selection for heteroscedastic Gaussian linear models with dependence and in particular in the long memory setting.

An other line of research concerns an extension of Theorem 2.1 for non linear models. Indeed, in the independent setting, a general model selection for non linear models is given in [Mas07] (Theorem 4.18). By combining a Gaussian concentration inequality together with a chaining argument for dependent variables, we believe that it is possible to generalize the ℓ_0 penalization strategy for non linear models.

Our work strongly relies on the Gaussian assumption. It would be also interesting to provide model selection results for non Gaussian noise. Note that [Gen14] gives a general model selection theorem for linear models, under moment conditions. It would be interesting to revisit these results in the context of long range dependence.

As illustrated in the last sections, it appears to be possible to adapt the slope heuristics for calibrating penalties in the context of regression with dependent errors. It would be more satisfactory to provide justification of the slope heuristics in this context. A first step would be to justify the slope heuristics for regression with short memory errors. Finally, note that model selection for density estimation under mixing conditions with resampling penalties has been studied in [Ler11]. This strategy is computationally expensive but it deserves to be investigated for regression under short and long memory errors.

Acknowledgment

The authors are grateful to Anne Philippe for helpful discussions and suggestions about statistics of long memory processes. We also thank the referees and the associate editor for their suggestions, which helped improve the first version of this paper.

7 Proofs

7.1 Proof of Theorem 2.1

We adapt the proof of Theorem 2.2 in [Gir14] in the framework of dependent Gaussian errors. Starting from the definition of $\hat{m}$, see Equation (2.1), we find that for all $m \in \mathcal{M}$

$$\|Y - \hat{\mu}_{\hat{m}}\|_n^2 + \operatorname{pen}(\hat{m}) \leq \|Y - \hat{\mu}_m\|_n^2 + \operatorname{pen}(m).$$

Next,

$$\|\varepsilon + (\mu^* - \hat{\mu}_{\hat{m}})\|_n^2 + \operatorname{pen}(\hat{m}) \leq \|\varepsilon + (\mu^* - \hat{\mu}_m)\|_n^2 + \operatorname{pen}(m),$$

and thus

$$\|\varepsilon\|_n^2 + \|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2 + 2\langle \varepsilon, \mu^* - \hat{\mu}_{\hat{m}}\rangle_n + \operatorname{pen}(\hat{m}) \leq \|\varepsilon\|_n^2 + \|\mu^* - \hat{\mu}_m\|_n^2 + 2\langle \varepsilon, \mu^* - \hat{\mu}_m\rangle_n + \operatorname{pen}(m),$$

where $\langle \cdot, \cdot \rangle_n$ is the normalized inner product in $\mathbb{R}^n$: $\langle \cdot, \cdot \rangle_n = \frac{1}{n}\langle \cdot, \cdot \rangle$. Now, since $\mathbb{E}(\varepsilon) = 0$, we see that

$$\mathbb{E}\left[\langle \varepsilon, \mu^* - \hat{\mu}_m\rangle_n\right] = -\mathbb{E}\left[\langle \varepsilon, \hat{\mu}_m\rangle_n\right] = -\mathbb{E}\left(\|\operatorname{Proj}_{S_m}(\varepsilon)\|_n^2\right) \leq 0\,,$$

and finally we obtain that

$$\mathbb{E}\,\|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2 \leq \mathbb{E}\,\|\mu^* - \hat{\mu}_m\|_n^2 + \operatorname{pen}(m) + 2\mathbb{E}\left(\langle \varepsilon, \hat{\mu}_{\hat{m}} - \mu^*\rangle_n - \operatorname{pen}(\hat{m})\right).$$

The theorem can be directly derived from the next result:

Proposition 7.1.1. *For the penalty defined by Equation (2.2), there exist some constants $a > 1$ and $L_K \geq 0$ that only depend on K, and a random variable Z satisfying $\mathbb{E}(Z) \leq L_K \frac{\rho(\Sigma)}{n}$, such that*

$$2\langle \varepsilon, \hat{\mu}_{\hat{m}} - \mu^* \rangle_n - \text{pen}(\hat{m}) \leq a^{-1} \|\hat{\mu}_{\hat{m}} - \mu^*\|_n^2 + Z.$$

According to the proposition, we find that

$$\mathbb{E}\left[\|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2\right] \leq \mathbb{E}\left[\|\mu^* - \hat{\mu}_m\|_n^2\right] + \text{pen}(m) + a^{-1}\mathbb{E}\left[\|\hat{\mu}_{\hat{m}} - \mu^*\|_n^2\right] + \mathbb{E}(Z)$$

and

$$\frac{a-1}{a}\mathbb{E}\left[\|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2\right] \leq \mathbb{E}\left[\|\mu^* - \hat{\mu}_m\|_n^2\right] + \text{pen}(m) + L_K \frac{\rho(\Sigma)}{n}.$$

Thus,

$$\mathbb{E}\left[\|\mu^* - \hat{\mu}_{\hat{m}}\|_n^2\right] \leq C_K \left(\mathbb{E}\left[\|\mu^* - \hat{\mu}_m\|_n^2\right] + \frac{\rho(\Sigma)}{n} + \text{pen}(m)\right),$$

where $C_K = \max\left(\frac{a}{a-1}, \frac{aL_K}{a-1}\right)$ and the proof of Theorem 2.1 is complete.

7.2 Proof of Proposition 7.1.1

We first recall a well known inequality from Cirel'son, Ibragimov et Sudakov [CIS76].

Theorem 7.1. *Let $F : (\mathbb{R}^n, \|\cdot\|) \to \mathbb{R}$ be a 1-Lipschitz function and η a random vector in $\mathbb{R}^n$ such that $\eta \sim \mathcal{N}_n(0, \sigma^2 Id)$ for some $\sigma > 0$. Then there exists a random variable ξ following an exponential distribution of parameter 1 such that*

$$F(\eta) \leq \mathbb{E}\left[F(\eta)\right] + \sigma\sqrt{2\xi}.$$

Note that the Lipschitz condition is expressed with respect to the (non-normalized) euclidean norm $\|\cdot\|$ in $\mathbb{R}^n$. We derive the following lemma for the projection of Gaussian random vectors.

Lemma 7.1. *Let Σ be a $n \times n$ symmetric semidefinite matrix and S a linear subspace of $\mathbb{R}^n$. Let ε be a Gaussian random vector such that $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$. Then there exists a random variable ξ following an exponential distribution of parameter 1 such that*

$$\|\text{Proj}_S(\varepsilon)\|_n \leq \mathbb{E}\|\text{Proj}_S(\varepsilon)\|_n + \sqrt{\frac{\rho(\Sigma)}{n}}\sqrt{2\xi}.$$

Proof. Let $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$, then ε satisfies $\varepsilon = \sqrt{\Sigma}\eta$ with $\eta \sim \mathcal{N}_n(0, Id)$. Let S be a linear subspace of $\mathbb{R}^n$. We then check that the function $\eta \to \left\|\text{Proj}_S(\sqrt{\Sigma}\eta)\right\|_n$ is a Lipschitz function

$$\begin{aligned}
\left\|\text{Proj}_S(\sqrt{\Sigma}x) - \text{Proj}_S(\sqrt{\Sigma}y)\right\|_n &\leq \left\|\sqrt{\Sigma}(x-y)\right\|_n \\
&\leq \rho(\sqrt{\Sigma})\|x-y\|_n \\
&\leq \sqrt{\rho(\Sigma)}\|x-y\|_n = \sqrt{\frac{\rho(\Sigma)}{n}}\|x-y\|.
\end{aligned}$$

By applying Theorem 7.1 to the function $\eta \to \left\|\text{Proj}_S(\sqrt{\Sigma}\eta)\right\|_n$, we find that

$$\left\|\text{Proj}_S(\sqrt{\Sigma}\eta)\right\|_n \leq \mathbb{E}\left\|\text{Proj}_S(\sqrt{\Sigma}\eta)\right\|_n + \sqrt{\frac{\rho(\Sigma)}{n}}\sqrt{2\xi}.$$

□

We are now in position to prove Proposition 7.1.1. Let $\bar{S}_m$ be the linear space spanned by S_m and μ^*. By applying the inequality $2\langle x, y\rangle_n \leq a\|x\|_n^2 + \|y\|_n^2/a$ for $a > 1$, we find that

$$\begin{aligned}
2\langle \varepsilon, \hat{\mu}_{\hat{m}} - \mu^* \rangle_n - \text{pen}(\hat{m}) &= 2\langle \text{Proj}_{\bar{S}_{\hat{m}}}(\varepsilon), \hat{\mu}_{\hat{m}} - \mu^* \rangle_n - \text{pen}(\hat{m}) \\
&\leq a\left\|\text{Proj}_{\bar{S}_{\hat{m}}}(\varepsilon)\right\|_n^2 + a^{-1}\|\hat{\mu}_{\hat{m}} - \mu^*\|_n^2 - \text{pen}(\hat{m}) \\
&\leq Z + a^{-1}\|\hat{\mu}_{\hat{m}} - \mu^*\|_n^2,
\end{aligned}$$

where

$$Z = a \left\| \mathrm{Proj}_{\bar{S}_{\hat{m}}}(\varepsilon) \right\|_n^2 - \mathrm{pen}(\hat{m}).$$

Now, we can write that

$$\begin{aligned}
\mathbb{E}(Z) = \mathbb{E}\left[a \left\| \mathrm{Proj}_{\bar{S}_{\hat{m}}}(\varepsilon) \right\|_n^2 - \mathrm{pen}(\hat{m}) \right] & \\
& \leq a \mathbb{E}\left[\max_{m \in \mathcal{M}} \left(\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 - \frac{1}{a} \mathrm{pen}(m) \right) \right] \\
& \leq a \sum_{m \in \mathcal{M}} \mathbb{E}\left[\left(\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 - \frac{1}{a} \mathrm{pen}(m) \right)_+ \right].
\end{aligned}$$

Let $m \in \mathcal{M}$. We start from the elementary inequality

$$\mathbb{E}\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n \quad \leq \quad \left(\mathbb{E}\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 \right)^{1/2}. \tag{7.1}$$

We can show that the quantity on the right side in (7.1) is exactly equal to $\sqrt{\frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{\bar{S}_m} \Sigma \right)}$ (see the arguments below). However $\bar{S}_m$ is unknown because it depends on μ^* and thus we can not directly define the penalty in function of $\mathrm{tr}\left(\mathrm{Proj}_{\bar{S}_m} \Sigma \right)$. We then use the decomposition

$$\mathrm{Proj}_{\bar{S}_m} = \mathrm{Proj}_{S_m} \oplus^{\perp} \mathrm{Proj}_{V_m},$$

where V_m is the orthogonal to S_m in $\bar{S}_m$. Note that the dimension of V_m is (at most) one. By Pythagoras theorem $\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 = \left\| \mathrm{Proj}_{S_m}(\varepsilon) \right\|_n^2 + \left\| \mathrm{Proj}_{V_m}(\varepsilon) \right\|_n^2$. Now

$$\begin{aligned}
\mathbb{E}\left\| \mathrm{Proj}_{S_m}(\varepsilon) \right\|_n^2 = \frac{1}{n} \mathrm{tr}\, \mathbb{E}\left(\varepsilon^t \mathrm{Proj}_{S_m} \varepsilon \right) & \\
= \frac{1}{n} \mathrm{tr}\, \mathbb{E}\left(\varepsilon \varepsilon^t \mathrm{Proj}_{S_m} \right) = \frac{1}{n} \mathrm{tr}\left(\Sigma \mathrm{Proj}_{S_m} \right) = \frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma \right),
\end{aligned}$$

and

$$\mathbb{E}\left\| \mathrm{Proj}_{V_m}(\varepsilon) \right\|_n^2 = \frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{V_m} \Sigma \right) \leq \frac{\rho(\Sigma)}{n}.$$

Finally

$$\mathbb{E}\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 \leq \frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma \right) + \frac{\rho(\Sigma)}{n}. \tag{7.2}$$

According to Lemma 7.1 and using the inequalities (7.1) and (7.2), there exists a random variable ξ_m following an exponential distribution of parameter 1 such that

$$\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n \leq \sqrt{\frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma \right) + \frac{\rho(\Sigma)}{n}} + \sqrt{\frac{\rho(\Sigma)}{n}} \sqrt{2\xi_m}.$$

Thus, the random variable Z satisfies

$$\begin{aligned}
\mathbb{E}(Z) \leq a \sum_{m \in \mathcal{M}} \mathbb{E}\left[\left(\left\| \mathrm{Proj}_{\bar{S}_m}(\varepsilon) \right\|_n^2 - \frac{1}{a} \mathrm{pen}(m) \right)_+ \right] & \\
\leq a \sum_{m \in \mathcal{M}} \mathbb{E}\Bigg[\Bigg(\left(\sqrt{\frac{1}{n} \mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma \right) + \frac{\rho(\Sigma)}{n}} + \sqrt{\frac{\rho(\Sigma)}{n}} \sqrt{2\xi_m} \right)^2 & \\
- \frac{1}{a} \mathrm{pen}(m) \Bigg)_+ \Bigg].
\end{aligned}$$

We assume as in (2.2) that

$$\mathrm{pen}(m) \geq \frac{K}{n} \left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma \right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m} \right)} \right)^2.$$

Then,

$$\mathbb{E}(Z) \leq \frac{a}{n} \sum_{m \in \mathcal{M}} \mathbb{E}\Bigg[\Bigg(\left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2\xi_m} \right)^2 - \frac{K}{a} \left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2 \Bigg)_+ \Bigg].$$

Using the inequality $(x+y)^2 \leq (1+\alpha)x^2 + (1+\alpha^{-1})y^2$, and taking $\alpha = \frac{K-a}{a}$ for $K > a > 1$, we find that

$$\begin{aligned}
&\left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2\xi_m} \right)^2 \\
&\qquad \leq \left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right. \\
&\qquad\qquad \left. + \sqrt{\rho(\Sigma)} \sqrt{2\left(\xi_m - \log\left(\frac{1}{\pi_m}\right)\right)_+} \right)^2 \\
&\qquad \leq \frac{K}{a} \left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2 \\
&\qquad\qquad + \frac{2K\rho(\Sigma)}{K-a} \left(\xi_m - \log\left(\frac{1}{\pi_m}\right)\right)_+ .
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2\xi_m} \right)^2 \\
&\qquad - \frac{K}{a} \left(\sqrt{\mathrm{tr}\left(\mathrm{Proj}_{S_m} \Sigma\right) + \rho(\Sigma)} + \sqrt{\rho(\Sigma)} \sqrt{2 \log\left(\frac{1}{\pi_m}\right)} \right)^2 \Bigg)_+ \Bigg] \\
&\qquad\qquad \leq \mathbb{E}\left[\frac{2K\rho(\Sigma)}{K-a} \left(\xi_m - \log\left(\frac{1}{\pi_m}\right)\right)_+ \right] \leq \frac{2K\rho(\Sigma)}{K-a} \pi_m,
\end{aligned}$$

because $\mathbb{E}\left[\left(\xi_m - \log\left(\frac{1}{\pi_m}\right)\right)_+ \right] = \exp(-\log(\frac{1}{\pi_m})) = \pi_m$. Since $\sum_{m \in \mathcal{M}} \pi_m = 1$, we finally obtain that

$$\mathbb{E}(Z) \leq a \sum_{m \in \mathcal{M}} \frac{2K}{K-a} \pi_m \frac{\rho(\Sigma)}{n} = \frac{2aK}{K-a} \frac{\rho(\Sigma)}{n} .$$

For any $K > 1$, take $a = \frac{K+1}{2}$. Then $K > a > 1$ is satisfied and the proof of Proposition 7.1.1 is complete with $L_K = \frac{2K^2+2K}{K-1}$.

7.3 Proof of Lemma 3.1

For any $m \in \{1, \ldots, n\}$ and any $j \in \{1, \ldots, m\}$, we define the discrete interval

$$I_j = \left\{ i \in \{1, \ldots, n\} : \frac{i}{n} \in \left[\frac{(j-1)}{m}, \frac{j}{m} \right[\right\} ,$$

and we denote by $\ell(j)$ the length of I_j: $\ell(j) = \mathrm{Card}(I_j)$. Note that, for all j, $[n/m] \leq \ell_j \leq [n/m] + 1$. The linear space S_m induced by the family of piecewise polynomials of degree at most r on the regular partition

of size m of the interval $[0,1]$ is the space generated by the $(r+1)m$ columns of the design

$$X = \begin{pmatrix}
1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
1 & 2 & \dots & 2^r & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \ell_1 & \dots & \ell_1^r & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 1 & 2 & \dots & 2^r & 0 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 1 & \ell_2 & \dots & \ell_2^r & 0 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 1 & 1 & \dots & 1 \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 1 & 2 & \dots & 2^r \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 1 & \ell_m & \dots & \ell_m^r
\end{pmatrix}.$$

Let c_k be the k-th column of the matrix X. Note that these columns are not all orthogonal, but they are linearly independent.

For $k \in \{1, \dots, m\}$, let V_k be the linear subspace of $\mathbb{R}^n$ generated by the c_j's for $j \in \{(k-1)(r+1)+1, \dots, k(r+1)\}$. Note that the subspaces V_k are orthogonal subspaces, so that

$$\left\|\mathrm{Proj}_{S_m}(\varepsilon)\right\|_n^2 = \sum_{k=1}^{m} \left\|\mathrm{Proj}_{V_k}(\varepsilon)\right\|_n^2 .$$

We shall prove that there exists a constant $C > 0$ such that, for any $k \in \{1, \dots, m\}$,

$$n\mathbb{E}\left(\left\|\mathrm{Proj}_{V_k}(\varepsilon)\right\|_n^2\right) \leq C\frac{n^{1-\gamma}}{m^{1-\gamma}} . \tag{7.3}$$

If (7.3) is true then the proof of Lemma 3.1 is easy to complete. Indeed

$$\mathrm{tr}\left(\mathrm{Proj}_{S_m}\Sigma\right) = n\mathbb{E}\left(\left\|\mathrm{Proj}_{S_m}(\varepsilon)\right\|_n^2\right) = \sum_{k=1}^{m} n\mathbb{E}\left(\left\|\mathrm{Proj}_{V_k}(\varepsilon)\right\|_n^2\right) \leq Cm^{\gamma}n^{1-\gamma} .$$

It remains to prove (7.3). In fact, it suffices to prove (7.3) for V_1, the argument being unchanged for the other V_k's. Let $e_k = c_k/\sqrt{c_k^t c_k}$, so that $n\|e_k\|_n^2 = 1$, and let X_1 be the $n \times (r+1)$ matrix composed of the $(r+1)$ columns $e_1, \dots e_{r+1}$. We can write

$$\mathrm{Proj}_{V_1}(\varepsilon) = \alpha_1 e_1 + \dots + \alpha_{r+1} e_{r+1},$$

where

$$(\alpha_1, \dots, \alpha_{r+1})^t = (X_1^t X_1)^{-1} X_1^t \varepsilon .$$

Clearly

$$\sqrt{\sum_{k=1}^{r+1} \alpha_k^2} \leq \rho\left((X_1^t X_1)^{-1}\right) \sqrt{\sum_{k=1}^{r+1} (e_k^t \varepsilon)^2} , \tag{7.4}$$

where $\rho\left((X_1^t X_1)^{-1}\right)$ is the spectral radius of $(X_1^t X_1)^{-1}$. Since

$$n\left\|\mathrm{Proj}_{V_1}(\varepsilon)\right\|_n^2 \leq (r+1)^2 \sum_{k=1}^{r+1} \alpha_k^2 ,$$

we infer from (7.4) that

$$n\mathbb{E}\left(\left\|\mathrm{Proj}_{V_1}(\varepsilon)\right\|_n^2\right) \leq \left((r+1)\rho\left((X_1^t X_1)^{-1}\right)\right)^2 \sum_{k=1}^{r+1} \mathbb{E}\left((e_k^t \varepsilon)^2\right) . \tag{7.5}$$

Before going further, we need to check that $\rho\left((X_1^tX_1)^{-1}\right)$ is uniformly bounded: indeed this quantity depends on the length ℓ_1, which can be as large as n. This is true, because $X_1^tX_1$ tends to A as $\ell_1 \to \infty$, where A is an invertible $(r+1)\times(r+1)$ matrix (in fact one can check that $A_{i,j} = \sqrt{(2j+1)(2i+1)}/(j+i+1)$). It follows that, as ℓ_1 varies, $\rho\left((X_1^tX_1)^{-1}\right)$ is a sequence of positive numbers converging to $\rho(A^{-1})$: it is therefore uniformly bounded. It follows from (7.5) that there exists $K>0$ such that

$$n\mathbb{E}\left(\left\|\mathrm{Proj}_{V_1}(\varepsilon)\right\|_n^2\right) \le K\sum_{k=1}^{r+1}\mathbb{E}\left((e_k^t\varepsilon)^2\right)\,.$$

Hence (7.3) will be proved for V_1 if there exists $C_1>0$ such that, for any $k\in\{1,\dots,r+1\}$,

$$\mathbb{E}\left((e_k^t\varepsilon)^2\right) = \mathbb{E}\left(\left(\frac{c_k^t\varepsilon}{\sqrt{c_k^tc_k}}\right)^2\right) \le C_1\frac{n^{1-\gamma}}{m^{1-\gamma}}\,. \tag{7.6}$$

It remains to prove (7.6). Let then $k\in\{1,\dots,r+1\}$. By stationarity,

$$\mathbb{E}\left(\left(c_k^t\varepsilon\right)^2\right) = \sum_{i=1}^{\ell_1}\sum_{j=1}^{\ell_1} i^kj^k\gamma_\varepsilon(j-i) \le \gamma_\varepsilon(0)\sum_{i=1}^{\ell_1} i^{2k} + 2\sum_{j=1}^{\ell_1}|\gamma_\varepsilon(j)|\sum_{i=1}^{\ell_1-j} i^k(i+j)^k\,.$$

Now, by Cauchy-Schwarz,

$$\sum_{i=1}^{\ell_1-j} i^k(i+j)^k \le \sum_{i=1}^{\ell_1} i^{2k} = c_k^tc_k\,.$$

Combining the two last inequalities, we get

$$\mathbb{E}\left(\left(\frac{c_k^t\varepsilon}{\sqrt{c_k^tc_k}}\right)^2\right) \le \gamma_\varepsilon(0) + 2\sum_{j=1}^{\ell_1}|\gamma_\varepsilon(j)|\,. \tag{7.7}$$

Now, recall that (3.4) holds, that is $|\gamma_\varepsilon(k)|\le\kappa(k+1)^{-\gamma}$ for some $\kappa>0$ and $\gamma\in(0,1)$. From (7.7), we easily infer that there exists $C_2>0$ such that

$$\mathbb{E}\left(\left(\frac{c_k^t\varepsilon}{\sqrt{c_k^tc_k}}\right)^2\right) \le C_2\ell_1^{1-\gamma}\,.$$

Since $[n/m]\le\ell_1\le[n/m]+1$, (7.6) easily follows. This completes the proof of Lemma 3.1.

7.4 Proof of Lemma 3.2

We keep the notations of the proof of Lemma 3.1. Recall that the case of regular regressograms corresponds to the degree $r=0$. In that case, the design matrix X of the proof of Lemma 3.1 contains only the m orthogonal columns filled with 0 and 1, and the linear space S_m has dimension m. Denote by $c_1,\dots,c_m$ the m columns of the design X.

We can write the exact expression of $\mathrm{Proj}_{S_m}(\varepsilon)$

$$\mathrm{Proj}_{S_m}(\varepsilon) = \bar\varepsilon_1c_1 + \bar\varepsilon_2c_2 + \cdots + \bar\varepsilon_mc_m\,, \quad \text{with} \quad \bar\varepsilon_k = \frac{1}{\ell_k}\sum_{i\in I_k}\varepsilon_i\,.$$

Consequently

$$n\left\|\mathrm{Proj}_{S_m}(\varepsilon)\right\|_n^2 = \ell_1\bar\varepsilon_1^2 + \ell_2\bar\varepsilon_2^2 + \cdots + \ell_m\bar\varepsilon_m^2\,.$$

Now, it follows from (3.5) that $\mathbb{E}(\bar\varepsilon_i^2)\le\kappa\ell_i^{-\gamma}$. Hence

$$\mathrm{tr}\left(\mathrm{Proj}_{S_m}\Sigma\right) = n\mathbb{E}\left(\left\|\mathrm{Proj}_{S_m}(\varepsilon)\right\|_n^2\right) \le \kappa\sum_{k=1}^m\ell_k^{1-\gamma}\,.$$

Since, for all j, $[n/m]\le\ell_j\le[n/m]+1$, we infer that there exists a positive constant C depending only on κ and γ such that

$$\mathrm{tr}\left(\mathrm{Proj}_{S_m}\Sigma\right) \le Cm^\gamma n^{1-\gamma}\,.$$

This concludes the proof of Lemma 3.2.

7.5 Proof of Inequality (2.4)

We give a short proof for the first inequality in (2.4). Let $(u_s)_{s=1,\dots,n}$ and $(\lambda_s)_{s=1,\dots,n}$ be an orthonormal basis of eigenvectors and a family of eigenvalues of Σ in decreasing order, in such a way that $\Sigma = \sum_{s=1}^{n} \lambda_s u_s u_s^t$. Let also $\sqrt{\Sigma} = \sum_{s=1}^{n} \sqrt{\lambda_s} u_s u_s^t$. For any vector space S of dimension d,

$$\begin{aligned} \operatorname{tr}\left(\operatorname{Proj}_S \Sigma\right) &= \operatorname{tr}\left(\operatorname{Proj}_S^t \operatorname{Proj}_S \sqrt{\Sigma}\sqrt{\Sigma}^t\right) \\ &= \operatorname{tr}\left(\operatorname{Proj}_S \sqrt{\Sigma}\sqrt{\Sigma}^t \operatorname{Proj}_S^t\right) \\ &= \|\operatorname{Proj}_S \sqrt{\Sigma}\|_F^2 \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. According to the Singular Value Decomposition (and more precisely to the Eckart–Young Theorem for low rank approximation, see [**?**]), this inertia term is maximized for S the linear space spanned by $(u_s)_{s=1\dots d}$ and thus

$$\operatorname{tr}\left(\operatorname{Proj}_S \Sigma\right) \leq \sum_{j=1}^{d} \lambda_j.$$

References

[Aka73] H. Akaike, Information theory and an extension of the maximum likelihood principle, Second International Symposium on Information Theory (Tsahkadsor, 1971), 1973, pp. 267–281.

[Arl19] Sylvain Arlot, Minimal penalties and the slope heuristics: a survey, arXiv preprint arXiv:1901.07277 (2019).

[Bar00] Yannick Baraud, Model selection for regression on a fixed design, Probability Theory and Related Fields **117** (2000), no. 4, 467–493.

[Bar02] ———, Model selection for regression on a random design, ESAIM: Probability and Statistics **6** (2002), 127–146.

[BCV01] Y Baraud, F Comte, and G Viennet, Adaptive estimation in autoregression or-mixing regression via model selection, The Annals of Statistics **29** (2001), no. 3, 839–875.

[Ber94] Jan Beran, Statistics for long-memory processes, Monographs on Statistics and Applied Probability, vol. 61, Chapman and Hall, New York, 1994.

[BF02] Jan Beran and Yuanhua Feng, Local polynomial fitting with long-memory, short-memory and antipersistent errors, Ann. Inst. Statist. Math. **54** (2002), no. 2, 291–311.

[BM01a] Lucien Birgé and Pascal Massart, Gaussian model selection, J. Eur. Math. Soc. (JEMS) **3** (2001), no. 3, 203–268.

[BM01b] Lucien Birgé and Pascal Massart, A generalized Cp criterion for gaussian model selection, Technical report, Universités de Paris 6 et Paris 7 (2001).

[BM07] ———, Minimal penalties for Gaussian model selection, Probability theory and related fields **138** (2007), no. 1-2, 33–73.

[BMM12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel, Slope heuristics: overview and implementation, Statistics and Computing **22** (2012), no. 2, 455–470.

[BS12] Jan Beran and Yevgen Shumeyko, On asymptotically optimal wavelet estimation of trend functions under long-range dependence, Bernoulli **18** (2012), no. 1, 137–176.

[CIS76] B. S. Cirel'son, I. A. Ibragimov, and V. N. Sudakov, Norms of Gaussian sample functions, Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975), 1976, pp. 20–41. Lecture Notes in Math., Vol. 550.

[CM95a] Sándor Csörgő and Jan Mielniczuk, Close short-range dependent sums and regression estimation, Acta Sci. Math. (Szeged) **60** (1995), no. 1-2, 177–196.

[CM95b] ———, Distant long-range dependent sums and regression estimation, Stochastic Process. Appl. **59** (1995), no. 1, 143–155.

[CM95c] ———, Nonparametric regression under long-range dependent normal errors, Ann. Statist. **23** (1995), no. 3, 1000–1014.

[DGM18] Jérôme Dedecker, Sébastien Gouëzel, and Florence Merlevède, Large and moderate deviations for bounded functions of slowly mixing Markov chains, Stoch. Dyn. **18** (2018), no. 2, 1850017, 38.

[DL93] Ronald A DeVore and George G Lorentz, Constructive approximation, vol. 303, Springer Science & Business Media, 1993.

[DMR94] Paul Doukhan, Pascal Massart, and Emmanuel Rio, The functional central limit theorem for strongly mixing processes, Ann. Inst. H. Poincaré Probab. Statist. **30** (1994), no. 1, 63–82.

[Gen08] Xavier Gendre, Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression, Electronic Journal of Statistics **2** (2008), 1345–1372.

[Gen14] ———, Model selection and estimation of a component in additive regression, ESAIM: Probability and Statistics **18** (2014), 77–116.

[Gir14] Christophe Giraud, Introduction to high-dimensional statistics, Chapman and Hall/CRC, 2014.

[HH90] Peter Hall and Jeffrey D. Hart, Nonparametric regression with long-range dependence, Stochastic Process. Appl. **36** (1990), no. 2, 339–351.

[HKP99] Peter Hall, Gérard Kerkyacharian, and Dominique Picard, On the minimax optimality of block thresholded wavelet estimators, Statist. Sinica **9** (1999), no. 1, 33–49.

[Hur51] Harold Edwin Hurst, Long-term storage capacity of reservoirs, Trans. Amer. Soc. Civil Eng. **116** (1951), 770–799.

[Joh99] Iain M. Johnstone, Wavelet shrinkage for correlated data and inverse problems: adaptivity results, Statist. Sinica **9** (1999), no. 1, 51–83.

[JS97] Iain M. Johnstone and Bernard W. Silverman, Wavelet threshold estimators for data with correlated noise, J. Roy. Statist. Soc. Ser. B **59** (1997), no. 2, 319–351.

[Ler11] Matthieu Lerasle, Optimal model selection for density estimation of stationary data under various mixing conditions, The Annals of Statistics **39** (2011), no. 4, 1852–1877.

[LX07] Linyuan Li and Yimin Xiao, On the minimax optimality of block thresholded wavelet estimators with long memory data, J. Statist. Plann. Inference **137** (2007), no. 9, 2850–2869.

[Mal73] Colin L Mallows, Some comments on Cp, Technometrics **15** (1973), no. 4, 661–675.

[Mas07] Pascal Massart, Concentration inequalities and model selection, Lecture Notes in Mathematics, vol. 1896, Springer, Berlin, 2007.

[MVN68] Benoit B. Mandelbrot and John W. Van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Rev. **10** (1968), 422–437.

[Rob97] P. M. Robinson, Large-sample inference for nonparametric regression with dependent errors, Ann. Statist. **25** (1997), no. 5, 2054–2083.

[Tou25] O Toussoun, Mémoire sur l'histoire du Nil. 3 vols, Cairo, L'Institut Français D'Archéologie Orientale (1925).

[TRYTV96] Lanh Tran, George Roussas, Sidney Yakowitz, and B. Truong Van, Fixed-design regression for linear time series, Ann. Statist. **24** (1996), no. 3, 975–991.

[Wan96] Yazhen Wang, Function estimation via wavelet shrinkage for long-memory data, The Annals of Statistics **24** (1996), no. 2, 466–484.

[Whi53] Peter Whittle, Estimation and information in stationary time series, Arkiv för matematik **2** (1953), no. 5, 423–434.