Annals of the Institute of Statistical Mathematics (2023) 75:567–592
https://doi.org/10.1007/s10463-022-00855-1

Data-driven model selection for same-realization predictions in autoregressive processes

Kare Kamila[1]

Received: 11 April 2021 / Revised: 13 September 2022 / Accepted: 11 October 2022 /
Published online: 27 November 2022

Abstract
This paper is about the one-step ahead prediction of the future of observations drawn from an infinite-order autoregressive AR(∞) process. It aims to design penalties (fully data driven) ensuring that the selected model verifies the efficiency property but in the non-asymptotic framework. We show that the excess risk of the selected estimator enjoys the best bias-variance trade-off over the considered collection. To achieve these results, we needed to overcome the dependence difficulties by following a classical approach which consists in restricting to a set where the empirical covariance matrix is equivalent to the theoretical one. We show that this event happens with probability larger than $1 - c_0/n^2$ with $c_0 > 0$. The proposed data-driven criteria are based on the minimization of the penalized criterion akin to the Mallows's C_p.

Keywords Model selection · Oracle inequality · Efficiency · Autoregressive process · Data driven

1 Introduction

Consider observations $(X_1, X_2, \dots, X_n)$ arising from a trajectory of the process

$$X_t = f^*\left((X_{t-i})_{i\in\mathbb{N}^*}\right) + \sigma\,\xi_t \quad \text{for any } t \in \mathbb{Z}, \tag{1}$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is a sequence of zero-mean independent identically distributed random variables (i.i.d.r.v) satisfying $\mathbb{E}(|\xi_0|^4) < \infty$ and $f^* : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is a measurable function and $\sigma > 0$ an unknown constant.

Kare Kamila has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754362.

✉ Kare Kamila
kamilakare@gmail.com

[1] SAMM, Université Paris 1, Panthéon-Sorbonne, 90, rue de Tolbiac 75634 Paris, France

The problem is to estimate the function f^* using these observations. Process (1) is a particular case of the general class of affine causal process studied in Doukhan and Wintenberger (2008) and Bardet and Wintenberger (2009).

The study of this type of process more often requires the classical regularity condition on the function f^*, which are not restrictive at all and remain valid in various time series models. This condition can be stated as follows:

$$\sum_{k=1}^{\infty} \left(\sup_{x \in \mathbb{R}^{\infty}} \left| \frac{\partial}{\partial x_k} f^*(x) \right| \right) < 1, \tag{2}$$

provided that that f^* admits partial derivatives on $\mathbb{R}^{\mathbb{N}}$. Under (2) and if the noise ξ_0 admits r-order moments, Doukhan and Wintenberger (2008) showed that there exists a stationary, mixing and ergodic solution to (1) admitting r-order moments.

Moreover, Bardet and Wintenberger (2009) studied the consistency and the asymptotic normality of the Quasi-maximum log-likelihood estimator (QMLE) of $\theta^* = (\theta_i^*)_{i \in \mathbb{N}}$ in the case $f^* = f_{\theta^*}$.

In this paper, we will focus only on processes with a linear regression function (f_{θ^*}) with respect to the past and depending on some parameter $\theta^* \in \mathbb{R}^{\mathbb{N}}$; that is

$$f^*(X_{t-1}, X_{t-2}, \ldots) = f_{\theta^*}(X_{t-1}, X_{t-2}, \ldots) = \sum_{i=1}^{\infty} \theta_i^* X_{t-i}. \tag{3}$$

For such processes, condition (2) is rewritten as

$$\mathbf{A1}: \qquad \sum_{i=1}^{\infty} |\theta_i^*| < 1.$$

Even if this condition reduces the set of parameters a bit, the class of AR(∞) processes checking the condition **A1** is rich and of practical importance because it contains all invertible causal ARMA(p, q) processes and it is very useful for prediction given the past. Moreover, contrary to the autocovariance of ARMA(p, q) processes which decays exponentially fast, AR(∞) are able to model more complex behavior such as slower decay of the covariance structure.

Henceforth, let observations $(X_1, X_2, \ldots, X_n)$ be a trajectory of the solution $X := (X_t)_{t \in \mathbb{Z}}$ of (1) verifying **A1**. The goal of this paper is to predict the next value X_{n+1}. In fact, if θ^* were known, a simple prediction of X_{n+1} could be $f_{\theta^*}(X_n, X_{n-1}, \ldots)$ setting $X_t = 0$ for all $t \leq 0$. However, θ^* is generally unknown and it is impossible to provide a direct estimator since its coordinates are infinite. It is classical to identify a 'good' finite-dimensional model based on the data which can be done by sieve estimation where only a finite number of $\{\theta_i^*\}_{i=1}^K$ is estimated and letting K grows as the sample size increases. A usual approach to this is model selection and the goal is to provide a model with the prediction error as small as the oracle's one.

This question has already been addressed in the literature. Shibata (1980) was the first to tackle this issue. He proved that Akaike criterion is *asymptotically efficient* in the sense that the selected model achieves a smaller one-step mean squared error of prediction when it is fitted to predict an independent realization of the same process.

Following Shibata's asymptotically setting, Ing and Wei (2003) and Ing and Wei (2005) extended this result for same realization predictions. Indeed, they argued that the Shibata's idea to fit the model to another independent realization is unrealistic since in practice we only have one data at hand. The common feature of these works is their asymptotic framework.

Meanwhile, there were several authors which study this question in non-asymptotic regime. Goldenshluger and Zeevi (2001) in the nonparametric framework, studied how well a Gaussian process admitting an AR(∞) representation can be approximated by a finite-order AR model.

In Baraud et al. (2001a) and (b), they analyzed similar question, but a little bit different as observations arise from an auto-regressive model of order k. They proved an oracle inequality under several conditions, for instance the compactly supported base of the regression function. Moreover, they assume that the process is β-mixing, which is usually admitted, but quite hard to verify in practice. For linear processes, the τ-mixing is more suitable since its coefficients can be easily computed (see Comte et al., 2008) and be bounded by a function of the model parameter θ^* (see Doukhan and Wintenberger, 2008). In this work, we do not assume any mixing property of the process since the condition **A1** implies the τ-mixing property (see Doukhan and Wintenberger, 2008) and we will see that the decreasing rate of τ-mixing coefficients is bounded by the decreasing rate of the coefficients $\theta^* = (\theta_i^*)_{i\in\mathbb{N}}$.

Based on the above and following a model selection approach, our purpose in this work is to design adaptive penalties in such a way that the selected model mimics the oracle when observations arise from AR(∞) under mild conditions, including the existence of the all order moments of the noise, the decreasing rate of the coefficients of $(\theta_i^*)_{i\in\mathbb{N}}$ so that thanks to a result by Doukhan and Wintenberger (2008), the generating process has nice properties such as stationarity, τ-mixing.

The main contribution of this paper is to prove that the excess risk of the selected estimator enjoys the best bias-variance trade-off over the considered collection.

The paper is organized as follows. The model selection approach along with preliminary results is described in Sect. 2. The main results are presented in Sect. 3. Finally, Sect. 4 contains the proofs.

2 Model selection approach and preliminary results

Before entering properly into the description of our approach, let us introduce some notations.

2.1 Notations

We will use the following norms:

- $< ., . >$ is the usual scalar product and if $x, y \in \mathbb{R}^n$, $< ., . >_n = \frac{1}{n}\sum_{i=1}^n x_i y_i$;
- The usual Euclidean norm on $\mathbb{R}^\nu$, with $\nu \geq 1$, is denoted by $\|.\|$ and its normalized version by $\|.\|_n$;

- $\|A\|_{\text{op}}$ is the operator norm of A as the square root of the largest eigenvalue of $A^\top A$. If A is symmetric, then $\|A\|_{\text{op}}$ is the largest (in absolute value) eigenvalue of A.
- If X is a $\mathbb{R}^\nu$-random variable and $r \geq 1$, we set $\|X\|_r = \left(\mathbb{E}\left[\|X\|^r\right]\right)^{1/r} \in [0, \infty]$, where $\mathbb{E}[Y]$ denotes the expected value of the random variable Y.

2.2 Model selection approach

Let S_m (shortly m) a model for f^* be the set of linear function f from $\mathbb{R}^{D_m}$ to $\mathbb{R}$ such that

$$f(x_1, x_2, \ldots, x_{D_m}) = f_\theta(x_1, x_2, \ldots, x_{D_m}) = \sum_{i=1}^{D_m} \theta_i x_i, \tag{4}$$

with $\theta = (\theta_1, \ldots, \theta_{D_m}) \in \Theta_m$ and Θ_m a compact set of $\mathbb{R}^{D_m}$ satisfying $\sup_{\theta \in \Theta_m} \sum_{i=1}^{D_m} |\theta_i| < 1$.

S_m can be viewed as an AR(D_m) model.

Given a predictor $f_\theta \in S_m$, its quality is measured by the quadratic loss

$$R(\theta) = \mathbb{E}\left[(X_{n+1} - f_\theta^{n+1})^2\right]$$

where $f_\theta^n = f_\theta(X_{n-1}, \ldots, X_{n-D_m})$. According to Bardet and Wintenberger (2009), the Bayes predictor which minimizes $R(\theta)$ over the set of all predictors is the inaccessible function f_{θ^*}. Let then introduce the excess loss of the predictor f_θ (with respect to f_{θ^*})

$$\ell(\theta, \theta^*) := R(\theta) - R(\theta^*) = \mathbb{E}\left[(f_{\theta^*}^{n+1} - f_\theta^{n+1})^2\right] \geq 0.$$

Given a model m, we define its best predictor $f_{\theta_m^*}$ by

$$\theta_m^* = \underset{\theta \in \Theta_m}{\operatorname{argmin}}\ R(\theta).$$

Its empirical version minimizing the least-squares contrast is

$$\widehat{\theta}_m = \underset{\theta \in \Theta_m}{\operatorname{argmin}}\ \gamma_n(\theta) \quad \text{where} \quad \gamma_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} (X_t - f_\theta^t)^2. \tag{5}$$

In this work, we will consider that the excess loss is measured on the design points, that is to say

$$\ell(\widehat{\theta}, \theta^*) = \mathbb{E}\left[\left\|F_{\widehat{\theta}} - F_{\theta^*}\right\|_n^2\right] \tag{6}$$

where $F_\theta := (f_\theta^1, \ldots, f_\theta^n)^\top$ and $\|x\|_n^2 = \frac{1}{n} \sum_{t=1}^n x_t^2$.

Given that all the models which can be considered must have finite dimensions for fixed n, making all S_m wrong models, it is classical to let the dimension of

competitive models grow with the number of observations. This will help reduce the excess loss and provide a better approximation of f_{θ^*}.

Let $\mathcal{M}_n$ a countable collection of hierarchical model S_m and K_n is the dimension of the largest model in $\mathcal{M}_n$ satisfying $|\mathcal{M}_n| \leq K_n < n$. We follow the classical approach of model selection which consists in minimizing the penalized LSE.

Let pen: $\mathcal{M}_n \to \mathbb{R}^+$ be a penalty function, possibly data-dependent, and define

$$\widehat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \{C(m)\} \qquad \text{with} \qquad C(m) := \gamma_n\big(\widehat{\theta}_m\big) + \operatorname{pen}(S_m). \tag{7}$$

The function pen can be a linear function of the model dimension (see for instance Birgé and Massart, (2001, 2007) among others) or a non linear one (see Lebarbier and Mary-Huard, 2004).

The best possible choice over $\mathcal{M}_n$ is m^* the so-called *oracle* defined as

$$m^* \in \arg \inf_{m \in \mathcal{M}_n} \ell(\widehat{\theta}_m, \theta^*). \tag{8}$$

The oracle m^* is unachievable since it depends on θ^* and the distribution $P_{(X_1,\ldots,X_n)}$ that are unknowns. However, we hope to select a model $\widehat{m}$ so that $\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)$ is closest to $\ell(\widehat{\theta}_{m^*}, \theta^*)$.

The goal of this paper is to propose a data-driven penalty in order to obtain an oracle inequality

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1 \inf_{m \in \mathcal{M}_n} \left\{\ell(\widehat{\theta}_m, \theta^*)\right\} + \frac{C_2}{n} \tag{9}$$

with the leading constant C_1 close to one and $C_2 > 0$. This goal could rather be to show that the excess risk of the selected estimator $\widehat{\theta}_{\widehat{m}}$ realizes the best bias-variance trade-off, which would make our penalty an ideal choice in terms of excess risk.

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C'_1 \inf_{m \in \mathcal{M}_n} \left\{ \ell(\theta^*_m, \theta^*) + \operatorname{pen}(S_m) \right\} + \frac{C'_2}{n} \tag{10}$$

with the leading constant $C'_1 = 1 + \delta$ with $\delta > 0$ (and close to 0) and $C'_2 > 0$.

That is to say that the selected model $\widehat{m}$ will be large enough to reduce its bias, but not too large to avoid high variance.

2.3 Preliminary results and assumptions

As we are in dependence setting, we are going to leverage the τ-mixing property of $(X_t)_{t \in \mathbb{Z}}$ in order to obtain some exponential inequalities. The τ-mixing coefficients are a measure of the dependence of the process and has been introduced by Dedecker and Prieur (2005). This will help us build 'independents' random vectors and apply classical exponential inequalities. Let then introduce some notations.

Let $(\Omega, \mathcal{C}, \mathbb{P})$ be a probability space, $\mathcal{M}$ a σ-subalgebra of $\mathcal{C}$ and Z a random variable with values in a Banach space $\big(E, \|.\|_E\big)$. Assume that $\mathbb{E}|Z| < \infty$ and define

$$\tau^{(p)}(\mathcal{M}, Z) = \Big\| \sup_{f\in\Lambda(E)} \Big\{ \Big| \int f(x)\mathbb{P}_{Z|\mathcal{M}}(dx) - \int f(x)\mathbb{P}_{Z}(dx) \Big| \Big\} \Big\|_p$$

where $\Lambda(E)$ is the set of 1-Lipschitz function, i.e., the functions f from $\big(E, \|.\|_E\big)$ to $\mathbb{R}$ such that $|f(x) - f(y)| \leq \|x - y\|_E$.

Using definition of τ, we will measure the dependence of the strictly stationary sequence $(Z_t)_{t\in\mathbb{Z}}$ thanks to the coefficients defined as follows. For any $s \geq 0$, let introduce the norm $\|x - y\|_{\mathbb{R}^k} = (|x_1 - y_1| + \cdots + |x_k - y_k|)$ and setting $\mathcal{M}_i = \sigma(Z_t, t \leq i)$ and if $\mathbb{E}(|Z_1|) < \infty$, let

$$\tau^{(p)}_{Z,\infty}(s) = \sup_{l>0} \Big\{ \max_{1\leq k\leq l} \frac{1}{k} \sup \Big\{ \tau^{(p)}\big(\mathcal{M}_i, (Z_{i_1}, \ldots, Z_{i_k})\big) \ , i + s \leq i_1 < \cdots < i_k \Big\} \Big\}.$$

Finally, the time series $(Z_t)_{t\in\mathbb{Z}}$ is $\tau^{(p)}_{Z,\infty}$-*weakly dependent* when its coefficients $\tau^{(p)}_{Z,\infty}$ tend to 0 as s tends to infinity.

Proposition 3 that is a consequence of Theorem 3.1 in Doukhan and Wintenberger (2008) gives a link between the τ-mixing coefficients of the process $(X_t)_{t\in\mathbb{Z}}$ and the coefficients θ^*_i of model (3).

As we are going to need independence for block of random variables, let denote for $t = 1, \ldots, n$ the random vector $\mathbf{X}_t := (X_{t-1}, \ldots, X_{t-K_n})^\top$ One can see that the process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is also mixing with $\tau^{(1)}_{\mathbf{X},\infty}$ upper bounded by $K_n\, \tau^{(1)}_{X,\infty}$ (see Lemma 1).

Now, we construct random variables approximating $\mathbf{X}_t$'s enjoying the independence by block property. Let s_n, q_n two integers such that $n = 2\, s_n\, q_n$. We are going to build $2\, s_n$ blocks of length q_n so that the even index blocks are independent and so the odd index blocks.

For $k = 0, \ldots, s_n - 1$ let denote by

$$A_k = \big(\mathbf{X}_{2kq_n+1}, \ldots, \mathbf{X}_{(2k+1)q_n}\big) \quad \text{and} \quad B_k = \big(\mathbf{X}_{(2k+1)q_n+1}, \ldots, \mathbf{X}_{(2k+2)q_n}\big).$$

Proposition 4 recalls a result from Lerasle (2011) that is a consequence of the coupling in Dedecker and Prieur (2005). It allows to have the block independence property.

To prove the oracle inequality, we will assume some constraints on the observations.

A2 X_t is sub-Gaussian with variance proxy $\sigma_0^2 > 0$ i.e.,

$$\mathbb{E}[e^{\lambda X_t}] \leq e^{\lambda^2 \sigma_0^2/2} \quad \text{for any } \lambda > 0.$$

Condition **A2** implies that the vector $Z_t^m = (X_{t-1}, \ldots, X_{t-D_m})^\top$ which will be prominent in the proofs, is sub-Gaussian with variance proxy $D_m\, \sigma_0^2$. Indeed for any $v \in \mathbb{R}^{D_m}$ such that $\|v\| = 1$,

$$\mathbb{E}\Big[\exp\left(\lambda\, v^\top Z_t^m\right)\Big] = \mathbb{E}\left[\prod_{i=1}^{D_m}\exp\left(\lambda\, v_i X_{t-i}\right)\right]$$

$$\&le \prod_{i=1}^{D_m}\left\|\exp\left(\lambda\, v_i X_{t-i}\right)\right\|_{D_m}$$

$$\leq \prod_{i=1}^{D_m}\exp\left(\lambda^2\, D_m\, \sigma_0^2\, v_i^2/2\right)$$

$$\leq e^{\frac{\lambda^2}{2} D_m \sigma_0^2},$$

where the Inequality follows from Hölder's Inequality.

The following assumption provides a sufficient condition to ensure the invertibility of both $\widehat{\Sigma}_m := \mathbf{M}_m^\top \mathbf{M}_m$ and $\Sigma_m := \mathbb{E}\big[\widehat{\Sigma}_m\big]$ where $\mathbf{M}_m = \big[X_{i-1}, \ldots, X_{i-D_m}\big]_{i=1}^n$.

A3: For any $f_\theta \in S_m, < \alpha, \partial_\theta f_\theta >= 0$ a.s. $\implies \alpha = 0$

This condition means that the columns of the matrix $\mathbf{M}_m$ are linearly independents.

We will also need to bound eigenvalues of the matrices Σ_m for any $m \in \mathcal{M}_n$. To do that, we will leverage the relation between the spectral density of the process and these eigenvalues. Let us denote by r, the covariance function $r(h) := \mathbb{E}[X_t X_{t+h}]$ for any integer h. Let also introduce the function $g : [-\pi, \pi[\longrightarrow \mathbb{C}$ such that for any λ,

$$g(\lambda) = \frac{1}{2\,\pi}\sum_{h\in\mathbb{Z}} r(h)\, e^{-ih\lambda},$$

which exists under **A1** with $|\theta_t^*| = O(t^{-\gamma})$ where $\gamma \geq 1$. Therefore, r is the inverse transform of g and $r(h) = \int_{-\pi}^{\pi} e^{ih\lambda}\, g(\lambda) d\lambda$ for any $h \in \mathbb{Z}$. We will assume that

A4: There exists a constant $a > 0$ such that $\inf_{-\pi\leq\lambda<\pi} g(\lambda) \geq a$.

This is a very weak assumption, and we are going to give the value of a for AR(p) process with $p \in \mathbb{N}$ and $p \geq 1$. Let denote $\theta^*(z) = 1 - \sum_{j=1}^p \theta_j^* z^j$, it is well known for such process that

$$g(\lambda) = \frac{\sigma^2}{2\,\pi\left|\theta^*(e^{-i\lambda})\right|^2}.$$

For instance for p equal to one, and $X_t = \theta_1^* X_{t-1} + \sigma\,\xi_t$ with $|\theta_1^*| < 1$, it follows

$$g(\lambda) = \frac{\sigma^2}{2\pi\left|1 - \theta_1^*\, e^{-i\lambda}\right|^2}$$

$$= \frac{\sigma^2}{2\pi\Big(1 - 2\,\theta_1^*\,\cos(\lambda) + (\theta_1^*)^2\Big)},$$

and then it is simple to see that

$$a := \frac{\sigma^2}{2\pi\,(1+|\theta_1^*|)^2} \leq \quad g(\lambda) \quad \leq \frac{\sigma^2}{2\pi\,(1-|\theta_1^*|)^2}.$$

For $p \geq 1$ and $X_t = \sum_{j=1}^{p} \theta_j^* X_{t-j} + \sigma\,\xi_t$ satisfying $\sum_{j=1}^{p} \theta_j^* < 1$ and $\theta_j^* \geq 0$, we have

$$\begin{aligned} g(\lambda) &= \frac{\sigma^2}{2\pi\left|1-\sum_{j=1}^{p}\theta_j^*\, e^{-ij\lambda}\right|^2} \\ &= \sigma^2\,(2\pi)^{-1}\left(1+\sum_{j=1}^{p}(\theta_j^*)^2 - 2\sum_{j=1}^{p}\theta_j^*\,\cos(j\lambda) + 2\sum_{k=1}^{p-1}\theta_k^*\Big\{\sum_{j=k+1}^{p}\theta_j^*\,\cos\big((j-k)\lambda\big)\Big\}\right)^{-1}. \end{aligned}$$

Thus, using $\big|\cos(x)\big| \leq 1$ for any real x, it follows for every λ that

$$\sigma^2\,(2\pi)^{-1}\left(1+\sum_{j=1}^{p}(\theta_j^*)^2 + 2\sum_{j=1}^{p}\theta_j^* + 2\sum_{k=1}^{p-1}\theta_k^*\Big\{\sum_{j=k+1}^{p}\theta_j^*\Big\}\right)^{-1} \leq \quad g(\lambda).$$

For such AR(p) process, one can take the constant a in **A4** to be equal to

$$a = \sigma^2\,(2\pi)^{-1}\left(1+\sum_{j=1}^{p}(\theta_j^*)^2 + 2\sum_{j=1}^{p}\theta_j^* + 2\sum_{k=1}^{p-1}\theta_k^*\Big\{\sum_{j=k+1}^{p}\theta_j^*\Big\}\right)^{-1}.$$

We can now state an important intermediate result which provides uniform lower and upper bound on the spectral norm of the matrices Σ_m.

Proposition 1 *Under* **A1** *with* $|\theta_t^*| = O(t^{-\gamma})$ *where* $\gamma \geq 2$, *we have for any* $m \in \mathcal{M}_n$

$$\big\|\Sigma_m\big\|_{\mathrm{op}} \leq \pi^{-1}\sum_{i=0}^{\infty}\big|\mathbb{E}[X_0\,X_i]\big| < \infty. \tag{11}$$

Moreover and under **A3-A4**, *it holds*

$$\big\|\Sigma_m^{-1}\big\|_{\mathrm{op}} \leq 1/a. \tag{12}$$

Let us introduce extra important notations. Let denote by μ the law of the vector $\mathbf{X}_t$ and

$$\Omega_n = \left\{\omega : \left|\frac{\|F_\theta\|_n^2}{\|F_\theta\|_\mu^2} - 1\right| \leq \frac{1}{2}, \quad \forall F_\theta \in \bigcup_{m,m'\in\mathcal{M}_n}(S_m + S_{m'})\right\}$$

where $\|F_\theta\|_\mu^2 := \frac{1}{n}\,\mathbb{E}\Big[\sum_{t=1}^{n}(f_\theta^t)^2\Big] = \int (f_\theta^1)^2 d\mu$. It is common to consider the set Ω_n which makes a link between the empirical norm $\|.\|_n$ and the to $\mathbb{L}_2$ norm (see for instance Baraud et al., 2001b; Hsu et al., 2011; Van de Geer, 2002, Comte & Genon-Catalot, 2020 among others). We will see that in our framework, Ω_n holds with large probability.

In all of this work, we assume that q_n was chosen to verify

$$\mathbf{A5}: \qquad \left(\frac{\log q_n}{q_n}\right)^{\gamma-1} \leq \frac{A}{n}, \tag{13}$$

for some constant A and $\gamma > 1$. Also we choose the integer s_n such that

$$\mathbf{A6}: \quad \frac{s_n}{2} \min\left\{\left(\frac{1}{2^7\,\sigma_0^2\,K_n}\right)^2, \frac{1}{2^8\,\sigma_0^2\,K_n}\right\} \geq 3\log n. \tag{14}$$

This means that s_n is of the form $s_n = C\,\log n$ where $C \geq 6\,\max\left\{\left(2^7\,\sigma_0^2\,K_n\right)^2, 2^8\,\sigma_0^2\,K_n\right\}$.

Proposition 2 *Under assumptions* **A1**, **A6** *and if* $|\theta_t^*| = O(t^{-\gamma})$ *with* $\gamma \geq 8$, *there exists a constant C such that*

$$\mathbb{P}(\Omega_n^c) \leq \frac{C}{n^3}. \tag{15}$$

3 Bias-variance result

We are now able to state the main result of the paper.

Theorem 1 *Let consider observations* $(X_1, \ldots, X_n)$ *arising from a solution of process* (1) *satisfying* **A1** *with* $|\theta_t^*| = O(t^{-\gamma})$ *where* $\gamma \geq 8$ *and also verifying* **A2** *and* **A4**. *Let* $\mathcal{M}_n$ *be some countable family of AR models satisfying* **A3** *and* **A5-A6**. *For any constant* $x > 2$, *let a penalty function pen*: $\mathcal{M}_n \to \mathbb{R}^+$ *such that*

$$\text{pen}(S_m) \geq 8\,x^3\sigma^2\,\frac{D_m}{n}. \tag{16}$$

Then, the LSE $\widehat{\theta}_{\widehat{m}}$ *with* $\widehat{m}$ *given in* (7), *satisfies*

$$\mathbb{E}\left[\left\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\right\|_n^2 \mathrm{I}1_{\Omega_n}\right] \leq C_1(x) \inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[\left\|F_{\theta^*} - F_{\theta_m^*}\right\|_n^2\right] + 2\,\text{pen}(S_m)\right\} + \frac{x(x+2)}{x-2}\,\frac{C_2}{n} \tag{17}$$

where $C_1(x) = \left(\frac{x+2}{x-2}\right)^2 > 1$ *and* $C_2 > 0$.

As we can see, this result is almost similar to that of Baraud et al. (2001b) obtained in nonparametric framework. However, their result is only valid if we want to estimate the function F_{θ^*} on some compact set. This restriction is lifted in our parametric framework.

4 Proofs

4.1 Proof of Theorem 1

Proof We follow the scheme of the proof of Baraud et al. (2001b). Let fix $m \in \mathcal{M}_n$. From definition (7), we have

$$\gamma_n\big(\widehat{\theta}_{\widehat{m}}\big) + \text{pen}(S_{\widehat{m}}) \leq \gamma_n\big(\widehat{\theta}_m\big) + \text{pen}(S_m) \leq \gamma_n(\theta^*_m) + \text{pen}(S_m). \tag{18}$$

Since,

$$\gamma_n\big(\widehat{\theta}_m\big) = \frac{1}{n}\sum_{t=1}^{n}(X_t - f^t_{\widehat{\theta}_m})^2 = \frac{1}{n}\sum_{t=1}^{n}(f^t_{\theta^*} - f^t_{\widehat{\theta}_m})^2 + \frac{\sigma^2}{n}\sum_{t=1}^{n}\xi_t^2 - \frac{2\,\sigma}{n}\sum_{t=1}^{n}\xi_t(f^t_{\widehat{\theta}_m} - f^t_{\theta^*}),$$

(18) yields to

$$\begin{aligned}\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\big\|_n^2 \leq \big\|F_{\theta^*_m} - F_{\theta^*}\big\|_n^2 \\ + \frac{2\,\sigma}{n}\sum_{t=1}^{n}\xi_t(f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*_m}) + \text{pen}(S_m) - \text{pen}(S_{\widehat{m}}).\end{aligned} \tag{19}$$

The difficult part of this proof is to handle the inner product $\frac{2}{n}\sum_{t=1}^{n}\sigma\,\xi_t(f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*_m})$, which can be rewritten as

$$\begin{aligned}\frac{2}{n}\sum_{t=1}^{n}\sigma\,\xi_t(f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*_m}) =& \frac{2}{n}\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_\mu \sum_{t=1}^{n}\sigma\,\xi_t\frac{(f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*_m})}{\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_\mu} \\ \leq& \frac{2}{n}\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_\mu \sup_{g\in B(\widehat{m},\mu)}\sum_{t=1}^{n}\sigma\,\xi_t\,g^t_\theta \\ \leq& x^{-1}\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_\mu^2 + n^{-2}\,x\Big(\sup_{g\in B(\widehat{m},\mu)}\sum_{t=1}^{n}\sigma\,\xi_t\,g^t_\theta\Big)^2\end{aligned}$$

since $2\,a\,b \leq x^{-1}\,a^2 + x\,b^2$ for any $x > 0$ and where

$$B(m',\mu) = \Big\{F_\theta \in S_m + S_{m'} \;:\; \|F_\theta\|_\mu^2 \leq 1\Big\}.$$

Moreover, on the set Ω_n, it holds

$$\begin{aligned}\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_\mu^2 \leq& 2\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*_m}\big\|_n^2 \\ \leq& 2\Big(\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\big\|_n + \big\|F_{\theta^*} - F_{\theta^*_m}\big\|_n\Big)^2 \\ \leq& 2\,(1+y)\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\big\|_n^2 + 2\,(1+y^{-1})\big\|F_{\theta^*} - F_{\theta^*_m}\big\|_n^2\end{aligned}$$

for some $y > 0$. As a result,

$$\frac{2}{n}\sum_{t=1}^{n}\sigma\,\xi_t(f^t_{\widehat{\theta}_{\widehat{m}}}-f^t_{\theta^*_m})\le 2\,\frac{(1+y)}{x}\left\|F_{\widehat{\theta}_{\widehat{m}}}-F_{\theta^*}\right\|_n^2+2\,\frac{(1+y^{-1})}{x}\left\|F_{\theta^*}-f_{\theta^*_m}\right\|_n^2$$
$$+\frac{x}{n^2}\Big(\sup_{g\in B(\widehat{m},\mu)}\sum_{t=1}^{n}\sigma\,\xi_t\,g^t_\theta\Big)^2.$$

Therefore, from (19), it holds on Ω_n

$$\Big(1-2\,\frac{(1+y)}{x}\Big)\left\|F_{\widehat{\theta}_{\widehat{m}}}-F_{\theta^*}\right\|_n^2\le\Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\left\|F_{\theta^*}-F_{\theta^*_m}\right\|_n^2$$
$$+\operatorname{pen}(S_m)-\operatorname{pen}(S_{\widehat{m}})+\frac{x\,\sigma^2}{n^2}\Big(\sup_{g\in B(\widehat{m},\mu)}\sum_{t=1}^{n}\xi_t\,g^t_\theta\Big)^2$$
$$\le\Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\left\|F_{\theta^*}-F_{\theta^*_m}\right\|_n^2+\operatorname{pen}(S_m)-\operatorname{pen}(S_{\widehat{m}})$$
$$+8\,x^3\sigma^2\,\frac{D_m+D_{\widehat{m}}}{n}+\frac{x\,\sigma^2}{n^2}\left[\Big(\sup_{g\in B(\widehat{m},\mu)}\sum_{t=1}^{n}\xi_t\,g^t_\theta\Big)^2-8n\,x^2D(S_{\widehat{m}})\right]_+$$

where $D(S_{\widehat{m}})=dim(S_m+S_{\widehat{m}})\le D_m+D_{\widehat{m}}$. Hence, using the condition on penalty (16),

$$\Big(1-2\,\frac{(1+y)}{x}\Big)\left\|F_{\widehat{\theta}_{\widehat{m}}}-F_{\theta^*}\right\|_n^2\le\Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\left\|F_{\theta^*}-F_{\theta^*_m}\right\|_n^2+2\operatorname{pen}(S_m)+x\,\sigma^2\,V_{\widehat{m}}\tag{20}$$

with

$$V_{m'}=\left[\Big(\sup_{g_\theta\in B(m',\mu)}\nu_n(g_\theta)\Big)^2-8\frac{x^2}{n}\,D(S_{m'})\right]_+,$$

where $\nu_n(g_\theta):=n^{-1}\sum_{t=1}^{n}\xi_t\,g^t_\theta$.

The proof will be established after controlling the expectation of $V_{\widehat{m}}$ which involves the supremum of an empirical process.

Now we leverage the mixing property in order to apply Talagrand's Inequality (Theorem 2) to tackle $\mathbb{E}[V_{\widehat{m}}]$.

We have

$$V_{\widehat{m}}=\left[\sup_{g_\theta\in B(\widehat{m},\mu)}\big(\nu_n(g_\theta)\big)^2-8\frac{x^2}{n}\,D(S_{\widehat{m}})\right]_+\le 2\sup_{g_\theta\in B(\widehat{m},\mu)}\Big(\nu_n(g_\theta)-\nu^*_n(g_\theta)\Big)^2+V^*_{\widehat{m}}\tag{21}$$

where

$$V^*_{\widehat{m}}=\left[2\sup_{g_\theta\in B(\widehat{m},\mu)}\big(\nu^*_n(g_\theta)\big)^2-8\frac{x^2}{n}\,D(S_{\widehat{m}})\right]_+\quad\text{and}\quad\nu^*_n(g_\theta)=\frac{1}{n}\sum_{t=1}^{n}\xi_t\,g_\theta(\mathbf{X}^*_t).$$

1/ **Control of** $\mathbb{E}\Big[\sup_{g_\theta \in B(\widehat{m},\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2\Big]$. Let $m' \in \mathcal{M}_n$ and $g_\theta \in B(m',\mu)$. Since the parameter set are compacts and $\theta \mapsto g_\theta$ is continuous, there exists $\theta_0 \in \Theta_m \cup \Theta_{m'}$ such that

$$\sup_{g_\theta \in B(m',\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2 = \frac{1}{n^2}\Big(\sum_{t=1}^{n} \xi_t \left(g_{\theta_0}(\mathbf{X}_t) - g_{\theta_0}(\mathbf{X}_t^*)\right)\Big)^2.$$

As ξ_t and $\mathcal{F}_t$ are independents, it follows that

$$\begin{aligned}\mathbb{E}\Big[\sup_{g_\theta \in B(m',\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2\Big] =& \frac{1}{n^2}\sum_{t=1}^{n} \mathbb{E}\Big[\xi_t^2 \left(g_{\theta_0}(\mathbf{X}_t) - g_{\theta_0}(\mathbf{X}_t^*)\right)^2\Big] \\ =& \frac{1}{n}\,\mathbb{E}\Big[\left(g_{\theta_0}(\mathbf{X}_0) - g_{\theta_0}(\mathbf{X}_0^*)\right)^2\Big]\end{aligned}$$

since $\mathbb{E}[\xi_0^2] = 1$. In addition,

$$\begin{aligned}\mathbb{E}\Big[\left(g_{\theta_0}(\mathbf{X}_0) - g_{\theta_0}(\mathbf{X}_0^*)\right)^2\Big] &= \sum_{i=1}^{D(S_{m'})}\sum_{j=1}^{D(S_{m'})} \theta_{0,i}\,\theta_{0,j}\,\mathbb{E}\big[(X_{-i} - X_{-i}^*)(X_{-j} - X_{-j}^*)\big] \\ &\le \Big(\sum_{i=1}^{D(S_{m'})} \theta_{0,i}\,\|X_{-i} - X_{-i}^*\|_2\Big)^2\end{aligned}$$

using Cauchy-Schwarz Inequality. It then follows as $\sum_{i=1}^{D(S_{m'})} |\theta_{0,i}| < 1$

$$\begin{aligned}\mathbb{E}\Big[\sup_{g_\theta \in B(m',\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2\Big] \le& \frac{1}{n}\left(\tau^{(2)}(q_n)\right)^2 \\ \le& \frac{C_\tau^2}{n}\left(\frac{\log q_n}{q_n}\right)^{2\gamma-2}\end{aligned}$$

where the last inequality follows from Proposition 3. Thus,

$$\begin{aligned}\mathbb{E}\Big[\sup_{g_\theta \in B(\widehat{m},\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2\Big] &\le \sum_{m' \in \mathcal{M}_n} \mathbb{E}\Big[\sup_{g_\theta \in B(m',\mu)} \Big(\nu_n(g_\theta) - \nu_n^*(g_\theta)\Big)^2\Big] \\ &\le K_n\,\frac{C_\tau^2}{n}\left(\frac{\log q_n}{q_n}\right)^{2\gamma-2} \\ &\le \frac{A^2\,C_\tau^2}{n^2},\end{aligned}$$

using Assumption **A5** and since $K_n \le n$.

2/ **Control of** $\mathbb{E}[V_{\widehat{m}}^*]$.

First, let us rewrite $\nu_n^*(g_\theta)$ for $g_\theta \in B(m',\mu)$. Setting $\mathbf{X}_t = (X_{t-1}^*, \ldots, X_{t-D(S_{m'})}^*)^\top$, we have

$$\begin{aligned} \nu_n^*(g_\theta) &= \frac{1}{n}\sum_{t=1}^{n} \xi_t\, g_\theta(\mathbf{X}_t^*) \\ &= \frac{1}{2\,s_n q_n}\sum_{k=0}^{s_n-1}\left(\sum_{i=1}^{q_n}\xi_{2kq_n+i}\, g_\theta(\mathbf{X}^*_{2kq_n+i}) + \sum_{i=1}^{q_n}\xi_{(2k+1)q_n+i}\, g_\theta(\mathbf{X}^*_{(2k+1)q_n+i})\right) \\ &= \nu^*_{n,1}(g_\theta) + \nu^*_{n,2}(g_\theta) \end{aligned}$$

with

$$\nu^*_{n,1}(g_\theta) = \frac{1}{s_n}\sum_{k=0}^{s_n-1}\nu^*_{n,1,k}(g_\theta) \quad \text{and} \quad \nu^*_{n,2}(g_\theta) = \frac{1}{s_n}\sum_{k=0}^{s_n-1}\nu^*_{n,2,k}(g_\theta)$$

where

$$\nu^*_{n,1,k}(g_\theta) = \frac{1}{2\,q_n}\sum_{i=1}^{q_n}\xi_{2kq_n+i}\, g_\theta(\mathbf{X}^*_{2kq_n+i}) \quad \text{and} \quad \nu^*_{n,2,k}(g_\theta) = \frac{1}{2\,q_n}\sum_{i=1}^{q_n}\xi_{(2k+1)q_n+i}\, g_\theta(\mathbf{X}^*_{(2k+1)q_n+i}).$$

Now let remark that $\nu^*_{n,1}(g_\theta)$ and $\nu^*_{n,2}(g_\theta)$ are both sum of s_n independent random variables by virtue of Proposition 4. Hence,

$$\begin{aligned} V^*_{\hat{m}} \le &\left(\sup_{g_\theta\in B(\hat{m},\mu)} 4\left(\nu^*_{n,1}(g_\theta)\right)^2 - 4\,x^2\,n^{-1}\,D(S_{\hat{m}})\right)_+ \\ &+\left(\sup_{g_\theta\in B(\hat{m},\mu)} 4\left(\nu^*_{n,2}(g_\theta)\right)^2 - 4\,x^2\,n^{-1}\,D(S_{\hat{m}})\right)_+. \end{aligned}$$

As a consequence it is sufficient to study $\mathbb{E}^*_1 := \mathbb{E}\left(\sup_{g\in B(\hat{m},\mu)} 4\left(\nu^*_{n,1}(g_\theta)\right)^2 - 4\,x^2\,n^{-1}\,D(S_{\hat{m}})\right)_+$ and the bound for $\mathbb{E}\left(\sup_{g_\theta\in B(\hat{m},\mu)} 4\left(\nu^*_{n,2}(g_\theta)\right)^2 - 4\,x^2\,n^{-1}\,D(S_{\hat{m}})\right)_+$ will follow by using analogous arguments.

Bounding $\mathbb{E}^*_1$

Since the noise (ξ_t) is not bounded, the process $\nu^*_{n,1}$ is not bounded either. Let's use the technique used in Comte and Genon-Catalot (2020) to overcome this difficulty. Therefore, we decompose ξ_t as

$$\xi_t = \eta_t + \epsilon_t, \quad \eta_t = \xi_t \mathrm{I1}_{|\xi_t|\le k_n},$$

where k_n is a deterministic sequence or a constant to be chosen later. We then have

$$\nu^*_{n,1}(g_\theta) = \upsilon^*_{n,1}(g_\theta) + \upsilon^*_{n,2}(g_\theta), \quad \text{where}$$

$$\upsilon^*_{n,1}(g_\theta) = \frac{1}{s_n}\sum_{k=0}^{s_n-1}\upsilon^*_{n,1,k}(g_\theta) \quad \text{with} \quad \upsilon^*_{n,1,k}(g_\theta) = \frac{1}{2\,q_n}\sum_{i=1}^{q_n}\eta_{2kq_n+i}\, g_\theta(\mathbf{X}^*_{2kq_n+i}) \quad \text{and}$$

$$v^*_{n,2}(g_\theta) = \frac{1}{s_n}\sum_{k=0}^{s_n-1} v^*_{n,2,k}(g_\theta) \quad \text{with} \quad v^*_{n,2,k}(g_\theta) = \frac{1}{2\,q_n}\sum_{i=1}^{q_n} \epsilon_{2kq_n+i}\, g_\theta(\mathbf{X}^*_{2kq_n+i}).$$

Thus,

$$\begin{aligned}\mathbb{E}^*_1 \leq & 8\,\mathbb{E}\left[\Big(\sup_{g_\theta\in B(\widehat{m},\mu)} \big(v^*_{n,1}(g_\theta)\big)^2 - 0.5\,x^2\,n^{-1}\,D(S_{\widehat{m}})\Big)_+\right] + 2\,\mathbb{E}\left[\sup_{g_\theta\in B(\widehat{m},\mu)} \big(v^*_{n,2}(g_\theta)\big)^2\right] \\ \leq & 8\sum_{m'\in\mathcal{M}_n} \mathbb{E}\left[\Big(\sup_{g_\theta\in B(m',\mu)} \big(v^*_{n,1}(g_\theta)\big)^2 - 0.5\,x^2\,n^{-1}\,D(S_{m'})\Big)_+\right] \qquad (22) \\ & + 2\,\mathbb{E}\left[\sup_{g_\theta\in B(\widehat{m},\mu)} \big(v^*_{n,2}(g_\theta)\big)^2\right]. \qquad (23)\end{aligned}$$

We start by bounding the term in (22). Let $m' \in \mathcal{M}_n$. In order to apply Theorem 2, one has to find M, H and v such that

$$\sup_{g_\theta\in B(m',\mu)} \big|v^*_{n,1,k}(g_\theta)\big| \leq M, \quad \mathbb{E}\Big[\sup_{g_\theta\in B(m',\mu)} \big|v_{n,1}(g_\theta)\big|^2\Big] \leq H^2,$$
$$\text{and} \quad \sup_{g\in B(m',\mu)} \mathrm{Var}\left(v^*_{n,1,k}(g_\theta))\right) \leq v.$$

• Since the noise is bounded here and from the assumption **A1**, the process (X_t) is also bounded. Indeed, under **A1**, there exists (ϕ^*_i) such that

$$X_t = \sum_{i=0}^{\infty} \phi^*_i\, \xi_{t-i} \quad \text{with} \quad \sum_{i=0}^{\infty} |\phi^*_i| < +\infty.$$

Therefore, $|X_t| \leq \Phi_0\, k_n$ with $\Phi_0 := \sum_{i=0}^{\infty} |\phi^*_i|$. Moreover, for any $g_\theta \in B(m', \mu)$, we have

$$\big|g_\theta(\mathbf{X}_t)\big| = \Big|\sum_{i=1}^{D(S_{m'})} \theta_i\, X_{t-i}\Big| \leq \Phi_0\, k_n \sum_{i=1}^{D(S_{m'})} |\theta_i| < \Phi_0\, k_n.$$

As a result, we have

$$\begin{aligned}\sup_{g_\theta\in B(m',\mu)} \big|v^*_{n,1,k}(g_\theta)\big| \leq & \frac{1}{2\,q_n} \sup_{g_\theta\in B(m',\mu)} \sum_{i=1}^{q_n} \big|\eta_{2kq_n+i}\, g_\theta(\mathbf{X}^*_{2kq_n+i})\big| \\ \leq & \frac{\Phi_0\, k_n^2}{2} := M.\end{aligned}$$

• Next, since the parameter set are compacts, there exists $\theta_0 \in \Theta_m \cup \Theta_{m'}$ such that

$$\sup_{g_\theta\in B(m',\mu)} \big|v^*_{n,1}(g_\theta)\big|^2 = \big|v^*_{n,1}(g_{\theta_0})\big|^2.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left|v^*_{n,1}(g_{\theta_0})\right|^2\right] &= \frac{1}{s_n}\,\mathbb{E}\left[\left|v^*_{n,1,0}(g_{\theta_0})\right|^2\right] \\ &= \frac{1}{4\,s_n\,q_n^2}\sum_{i,j=1}^{q_n}\mathbb{E}\left[\eta_i\,g_\theta(\mathbf{X}^*_i)\,\eta_j\,g_\theta(\mathbf{X}^*_j)\right] \\ &= \frac{1}{4\,s_n\,q_n^2}\sum_{i=1}^{q_n}\mathbb{E}\left[\left(\eta_i\,g_\theta(\mathbf{X}^*_i)\right)^2\right] \\ &\le \frac{\Phi_0^2\,k_n^2}{2\,n} \quad \le \quad \frac{\Phi_0^2\,k_n^2}{2\,n}\,D(S_{m'}) := H^2\end{aligned}$$

since $D(S_{m'}) \ge 1$.

- Lastly, as $\mathrm{Var}\,[X] \le E[X^2]$, it follows from the previous series of equations

$$\mathrm{Var}\left(v^*_{n,1,0}(g_\theta))\right) \le \mathbb{E}\left[\left|v^*_{n,1,0}(g_{\theta_0})\right|^2\right] \le \frac{\Phi_0^2\,k_n^2}{4\,q_n} := v.$$

As a consequence from Theorem 2 and taking $\alpha = \frac{1}{2}(\frac{x^2}{2\Phi_0^2\,k_n^2} - 1) > 0$, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\sup_{g_\theta\in B(m',\mu)}\left(v^*_{n,1}(g_\theta)\right)^2 - 0.5\,x^2\,n^{-1}\,D(S_{m'})\right)_+\right] \\ &\le \frac{2}{K}\left(\frac{\Phi_0^2\,k_n^2}{4q_n}e^{-Kq_nD(S_{m'})\left(\frac{x^2}{2\Phi_0^2\,k_n^2}-1\right)} + \frac{49\Phi_0^2k_n^4}{4n^2KC^2(\alpha)}e^{-2\sqrt{2}KC(\alpha)\frac{\sqrt{n}\sqrt{D(S_{m'})}}{k_n}}\right).\end{aligned}$$

Hence, there exists a constant K' such that

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{g_\theta\in B(m',\mu)}\left(v^*_{n,1}(g_\theta)\right)^2 - 0.5\,x^2\,n^{-1}\,D(S_{m'})\right)_+\right] \le \frac{K'}{n}. \tag{24}$$

- Now, let us upper bound the term in (23). For any $m' \in \mathcal{M}_n$ and any $g_\theta \in B(m', \mu)$, we have

$$g_\theta(\mathbf{X}_t) = \sum_{i=1}^{D(S_{m'})}\theta_i\,X_{t-i} \le \sup_{t-D(S_{m'})\le i<t}|X_i|\left(\sum_{i=1}^{D(S_{m'})}|\theta_i|\right) < \sup_{t-D(S_{m'})\le i<t}|X_i|.$$

Therefore,

$$\begin{aligned}v^*_{n,2,k}(g_\theta) =& \frac{1}{2\,q_n}\sum_{i=1}^{q_n}\xi_{2kq_n+i}\,g_\theta(\mathbf{X}^*_{2kq_n+i}) \\ <& \frac{1}{2\,q_n}\sum_{i=1}^{q_n}|\xi_{2kq_n+i}|\sup_{2kq_n+i-K_n\le t<2kq_n+i}|X^*_t| := Y^*_k,\end{aligned} \tag{25}$$

so that

$$\sup_{g\in B(\widehat{m},\mu)} \upsilon^*_{n,2}(g_\theta) < \frac{1}{s_n}\sum_{k=0}^{s_n-1} Y^*_k.$$

Let us notice that $(Y^*_k)_k$ is a family of independent random variables as $(\upsilon^*_{n,1,k}(g))_k$. Thus, it follows

$$\begin{aligned}\mathbb{E}\Big[\sup_{g_\theta\in B(\widehat{m},\mu)} \big|\upsilon^*_{n,2}(g_\theta)\big|^2\Big] &< \frac{1}{s_n^2}\sum_{i,j=0}^{s_n-1}\mathbb{E}\big[Y^*_i\,Y^*_j\big]\\ &< \frac{1}{s_n^2}\sum_{i=0}^{s_n-1}\mathbb{E}\big[Y^{*2}_i\big]\\ &= \frac{1}{s_n}\mathbb{E}\big[Y^{*2}_0\big].\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\big[Y^{*2}_0\big] &= \frac{1}{4\,q_n^2}\sum_{i,j=1}^{q_n}\mathbb{E}\Big[|\xi_i|\sup_{i-K_n\le t<i}|X^*_t|\,|\xi_j|\sup_{j-K_n\le t<j}|X^*_t|\Big]\\ &= \frac{1}{4\,q_n^2}\sum_{i=1}^{q_n}\mathbb{E}\Big[\Big(\sup_{i-K_n\le t<i}|X^*_t|\Big)^2\Big]\\ &= \frac{\mu_2}{4\,q_n},\end{aligned} \tag{26}$$

where $\mu_2 = \mathbb{E}[X_t^2] < \infty$. It follows

$$\mathbb{E}\Big[\sup_{g_\theta\in B(\widehat{m},\mu)} \big|\nu^*_{n,1}(g_\theta)\big|^2\Big] < \frac{\mu_2}{4\,s_n\,q_n} = \frac{\mu_2}{2\,n}. \tag{27}$$

Inequality (24) along with (27) yields to

$$\mathbb{E}^*_1 \le \frac{8\,K'}{n} + \frac{\mu_2}{n}.$$

We conclude that there exists $K > 0$

$$\mathbb{E}[V_{\widehat{m}}] \le \frac{K}{n}. \tag{28}$$

Returning to (20), and taking expectation on both sides, it then follows

$$\begin{aligned}&\Big(1-2\,\frac{(1+y)}{x}\Big)\mathbb{E}\Big[\big\|f^t_{\widehat{\theta}_{\widehat{m}}}-f^t_{\theta^*}\big\|_n^2\Big] \le \Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\mathbb{E}\Big[\big\|f^t_{\theta^*}-f^t_{\theta^*_m}\big\|_n^2\Big]\\ &+2\,\mathrm{pen}(S_m)+x\,\frac{K}{n}.\end{aligned} \tag{29}$$

For $y = \frac{x-2}{x+2} > 0$, so that $1+y = \frac{2x}{x+2}$ and $1+y^{-1} = \frac{2x}{x-2}$, we obtain

$$\mathbb{E}\left[\left\|f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*}\right\|_n^2\right] \leq C(x)\left(\mathbb{E}\left[\left\|f^t_{\theta^*} - f^t_{\theta^*_m}\right\|_n^2\right] + 2\,\mathrm{pen}(S_m)\right) + \frac{x(x+2)}{x-2}\,\frac{K}{n}$$

with $C(x) = \frac{(x+2)^2}{(x-2)^2} > 1$. □

4.2 Proof of Proposition 2

Since the collection $\mathcal{M}_n$ is hierarchical, we have

$$\begin{aligned}\mathbb{P}(\Omega_n^c) &\leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\exists F_\theta \in S_m : \left|\frac{\|f_\theta\|_n^2}{\|f_\theta\|_\mu^2} - 1\right| > \frac{1}{2}\right)\\ &\leq \sum_{m\in\mathcal{M}_n} \mathbb{P}(\Omega_m^c)\end{aligned}$$

where

$$\Omega_m = \left\{\left|\frac{\|F_\theta\|_n^2}{\|F_\theta\|_\mu^2} - 1\right| \leq \frac{1}{2} \quad \forall F_\theta \in S_m\right\}.$$

Let $m \in \mathcal{M}_n$. We have

$$\mathbb{P}(\Omega_m^c) \leq \mathbb{P}\left(\sup_{F_\theta\in S_m}\left|\frac{\|F_\theta\|_n^2}{\|F_\theta\|_\mu^2} - 1\right| > \frac{1}{2}\right).$$

Moreover,

$$\sup_{F_\theta\in S_m, \|F_\theta\|_\mu^2=1}\left|\frac{\|F_\theta\|_n^2}{\|F_\theta\|_\mu^2} - 1\right| > \frac{1}{2} \iff \sup_{F_\theta\in S_m, \|F_\theta\|_\mu^2=1}\left|\nu_n(F_\theta^2)\right| > \frac{1}{2}$$

with $\nu_n(F_\theta^2) = n^{-1}\sum_{t=1}^n \left((f_\theta^t)^2 - \mathbb{E}\left[(f_\theta^1)^2\right]\right)$. Hence,

$$\mathbb{P}\left(\sup_{f\in S_m}\left|\frac{\|F_\theta\|_n^2}{\|F_\theta\|_\mu^2} - 1\right| > \frac{1}{2}\right) \leq \mathbb{P}\left(\sup_{F_\theta\in S_m}\left|\nu_n(F_\theta^2)\right| > \frac{1}{2}\right).$$

For any $F_\theta \in S_m$, using the linearity we can write

$$\left(f_\theta(\mathbf{X}_t)\right)^2 = \left(\sum_{i=1}^{D_m}\theta_i X_{t-i}\right)^2 = \sum_{i,j=1}^{D_m}\theta_i\theta_j X_{t-i}X_{t-j} = \theta^\top\widehat{\Sigma}_{m,t}\theta$$

where $\theta = (\theta_1, \ldots, \theta_{D_m})^\top$, $\widehat{\Sigma}_{m,t} = Z_t^m(Z_t^m)^\top$ with $Z_t^m = (X_{t-1}, \ldots, X_{t-D_m})^\top$. So that with $\Sigma_m = \mathbb{E}\left[\widehat{\Sigma}_{m,t}\right]$, it follows

$$\nu_n^*(F_\theta^2) = \frac{1}{n}\sum_{t=1}^{n} \theta^\top \left(\widehat{\Sigma}_t - \Sigma\right)\theta = \theta^\top(\widehat{\Sigma}_m - \Sigma_m)\theta,$$

where $\widehat{\Sigma} = n^{-1}\sum_{t=1}^{n}\widehat{\Sigma}_t$. As a result,

$$\sup_{F_\theta \in S_m, \|F_\theta\|_\mu^2 = 1} |\nu_n^*(F_\theta^2)| \le \left\|\widehat{\Sigma}_m - \Sigma_m\right\|_{\text{op}}.$$

Indeed,

$$\begin{aligned}\sup_{F_\theta \in S_m} |\nu_n^*(F_\theta^2)| = \sup_{\theta:\, \sum|\theta_i|<1} \theta^\top(\widehat{\Sigma}_m - \Sigma_m)\theta &= \sup_{\theta:\, \sum|\theta_i|<1} \|\theta\|^2 \ \frac{\theta^\top(\widehat{\Sigma}_m - \Sigma_m)\theta}{\|\theta\|^2} \\ &\le \sup_{\theta:\, \|\theta\|^2 \le 1} \frac{\theta^\top(\widehat{\Sigma}_m - \Sigma_m)\theta}{\|\theta\|^2} = \left\|\widehat{\Sigma}_m - \Sigma_m\right\|_{\text{op}}\end{aligned}$$

since $-1 < \theta_i < 1$ ensures that $\|\theta\|^2 \le \sum|\theta_i|$. Hence,

$$\begin{aligned}\mathbb{P}\left(\sup_{F_\theta \in S_m, \|F_\theta\|_\mu^2 = 1} |\nu_n(F_\theta^2)| > \frac{1}{2}\right) &\le \mathbb{P}\left(\left\|\widehat{\Sigma}_m - \Sigma_m\right\|_{\text{op}} > \frac{1}{2}\right) \\ &\le \mathbb{P}\left(\left\|\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\right\|_{\text{op}} > \frac{1}{4}\right) + \mathbb{P}\left(\left\|\widehat{\Sigma}_m^* - \Sigma_m\right\|_{\text{op}} > \frac{1}{4}\right) \\ &=: \mathbb{P}_1 + \mathbb{P}_2.\end{aligned}$$

Using Lemma 3 with $u = 1/4$ and by virtue of **A6**, it follows

$$\begin{aligned}\mathbb{P}_2 &\le 2\ \exp\left(-3\log n\right) \\ &\le \frac{2}{n^3}.\end{aligned}$$

Now let bound $\mathbb{P}_1$. We know that for a $D_m \times D_m$ matrix A

$$\|A\|_{\text{op}} \le \|A\|_\infty := \max_{1\le i\le D_m} \sum_{j=1}^{D_m} |A_{ij}|.$$

Thus, from Markov's Inequality,

$$\begin{aligned}
\mathbb{P}_1 \leq & 4\,\mathbb{E}\left[\left\|\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\right\|_{\mathrm{op}}\right] \\
\leq & 4\,\mathbb{E}\left[\max_{1\leq i\leq D_m}\sum_{j=1}^{D_m}\left|\left(\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\right)_{i,j}\right|\right] \\
\leq & 4\sum_{j=1}^{D_m}\mathbb{E}\left[\left|\left(\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\right)_{i_0,j}\right|\right] \\
\leq & 4\sum_{j=1}^{D_m}\mathbb{E}\left[\left|X_{t-i_0}X_{t-j} - X^*_{t-i_0}X^*_{t-j}\right|\right].
\end{aligned}$$

Moreover, $\left|X_{t-i}X_{t-j} - X^*_{t-i}X^*_{t-j}\right| \leq \left|X_{t-i}\right|\left|X_{t-j} - X^*_{t-j}\right| + \left|X^*_{t-j}\right|\left|X_{t-i} - X^*_{t-i}\right|$ so that with Cauchy–Schwarz's Inequality,

$$\begin{aligned}
\mathbb{E}\left[\left|X_{t-i}X_{t-j} - X^*_{t-i}X^*_{t-j}\right|\right] \leq & 2\left\|X_0\right\|_2\left\|X_{t-1} - X^*_{t-1}\right\|_2 \\
\leq & 2\left\|X_0\right\|_2\tau^{(2)}(q_n).
\end{aligned}$$

Hence, using Proposition 3, it follows

$$\begin{aligned}
\mathbb{P}_1 \leq & 8\left\|X_0\right\|_2 D_m\,\tau^{(2)}(q_n) \\
\leq & 8\left\|X_0\right\|_2 D_m\; C_\tau\left(\frac{\log q_n}{q_n}\right)^{\gamma-1}.
\end{aligned}$$

Moreover, since $\gamma \geq 8$ and from assumption **A5**, one can find some constant A' such that

$$\left(\frac{\log q_n}{q_n}\right)^{\gamma-1} \leq \frac{A'}{n^4}.$$

As a result, with $c_0 := 8\left\|X_0\right\|_2 C_\tau A'$, it holds

$$\mathbb{P}_1 \leq \frac{c_0}{n^3}.$$

As a consequence,

$$\mathbb{P}(\Omega_n^c) \quad \leq \quad \frac{2 + c_0}{n^3}.$$

□

4.3 Proof of Proposition 1

Proof The proof of the will be based on the relation between the spectral density function and the maximum eigenvalues of the variance covariance matrix.

Denote by $u \in \mathbb{R}^{D_m}$ the normalized eigenvector associated with the largest eigenvalue $\lambda_{\max}(\Sigma_m)$. Hence,

$$\lambda_{\max}(\Sigma_m) = u^\top \Sigma_m\, u = \sum_{j,k=1}^{D_m} u_j\, r(j-k)\, u_k = \int_{-\pi}^{\pi} g(\lambda) \sum_{j,k=1}^{D_m} u_j\, e^{i(j-k)\lambda}\, u_k d\lambda$$
$$= \int_{-\pi}^{\pi} g(\lambda) \left| \sum_{j=1}^{D_m} u_j\, e^{ij\lambda} \right|^2 d\lambda \;\leq\; \sup_{-\pi\leq\lambda<\pi} g(\lambda) \int_{-\pi}^{\pi} \left| \sum_{j=1}^{D_m} u_j\, e^{ij\lambda} \right|^2 d\lambda$$
$$\leq \sup_{-\pi\leq\lambda<\pi} g(\lambda),$$

since, using Parseval identity, $\int_{-\pi}^{\pi} \left| \sum_{j=1}^{D_m} u_j\, e^{ij\lambda} \right|^2 d\lambda = \sum_{j=1}^{D_m} u_j^2 = 1$.

But, from Lemma 2 and since $\gamma \geq 2$, it follows

$$\left| \sup_{-\pi\leq\lambda<\pi} g(\lambda) \right| \leq \frac{1}{2\,\pi} \sum_{h\in\mathbb{Z}} |r(h)|$$
$$\leq \frac{C}{\pi} \sum_{h=0}^{+\infty} \frac{1}{(h+1)^\gamma} < \infty.$$

Given that Σ_m is symmetric, it follows

$$\left\|\Sigma_m\right\|_{\text{op}} = \lambda_{\max}(\Sigma_m) \leq \frac{C}{\pi} \sum_{h=0}^{+\infty} \frac{1}{(h+1)^\gamma},$$

which concludes the proof of (11).

Now we end by the proof of (12). Reasoning as above, and by virtue of **A4**, one can show that

$$\lambda_{\min}(\Sigma_m) \;\geq\; \inf_{-\pi\leq\lambda<\pi} g(\lambda) \geq a$$

which yields to

$$\left\|\Sigma_m^{-1}\right\|_{\text{op}} = \frac{1}{\lambda_{\min}(\Sigma_m)} \leq \frac{1}{a},$$

so that (12) is established.

□

4.4 Technical lemmas

Lemma 1 *Assume* **A1** *holds and* (X_t) *the mixing stationary solution of* (1). *Then, the process* $(\mathbf{X}_t)$ *is mixing and*

$$\tau^{(1)}_{\mathbf{X},\infty}(r) \leq K_n\, \tau^{(1)}_{X,\infty}(r-1). \tag{30}$$

Proof Let set by $\mathcal{M}^i_{\mathbf{X}} = \sigma(\mathbf{X}_t, t \le i)$ and $\mathcal{M}^i_X = \sigma(X_t, t \le i)$ for an integer i. One would like to bound $\tau^{(1)}\big(\mathcal{M}^i_{\mathbf{X}}, \big(\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k}\big)\big)$ for $j_k > \dots > j_1 \ge i + r$.

Let assume that the universe Ω is rich enough so that, one can find $\mathbf{X}^*_{j_l} = \big(X^*_{j_l-1}, \dots, X^*_{j_l-K_n}\big)^\top$ with $l = 1, \dots, k$ verifying

1. $\big(\mathbf{X}^*_{j_1}, \dots, \mathbf{X}^*_{j_k}\big)$ is distributed as $\big(\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k}\big)$ and independent of $\mathcal{M}^i_{\mathbf{X}}$;
2. $\big(X^*_{j_1-1}, \dots, X^*_{j_k-1}\big)^\top$ is distributed as $\big(X_{j_1-1}, \dots, X_{j_k-1}\big)^\top$ and independent of $\mathcal{M}^i_X$.

As a result,

$$\begin{aligned}
\tau^{(1)}\big(\mathcal{M}^i_{\mathbf{X}}, \big(\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k}\big)\big) \le & \sum_{l=1}^{k} \|\mathbf{X}_{j_l} - \mathbf{X}^*_{j_l}\|_1 = \sum_{l=1}^{k} \sum_{t=1}^{K_n} \mathbb{E}\big[|X_{j_l-t} - X^*_{j_l-t}|\big] \\
\le & K_n \sum_{l=1}^{k} \mathbb{E}\big[|X_{j_l-1} - X^*_{j_l-1}|\big] \\
= & K_n \left\| \big(X_{j_1-1}, \dots, X_{j_k-1}\big)^\top - \big(X^*_{j_1-1}, \dots, X^*_{j_k-1}\big)^\top \right\|_1 \\
= & K_n \, \tau^{(1)}\big(\mathcal{M}^i_X, \big(X_{j_1-1}, \dots, X_{j_k-1}\big)\big).
\end{aligned}$$

This fact along with the definition of $\tau^{(1)}_{\mathbf{X},\infty}(r)$ leads to (30).

□

Lemma 2 *Under* **A1** *with* $|\theta^*_t| = O(t^{-\gamma})$ *where* $\gamma > 1$, *we have*

$$r(h) = \mathbb{E}[X_0 X_h] = O\big((h+1)^{-\gamma}\big)$$

Proof By virtue of **A1**, the process $(X_t)_t$ is causal; that is there exists $(\phi_i)_{i\in\mathbb{N}}$ such that $X_t = \sum_{i=0}^{+\infty} \phi_i \, \xi_{t-i}$ with $\sum_{i=0}^{+\infty} |\phi_i| < \infty$. The sequence $(\phi_i)_{i\in\mathbb{N}}$ is given by the relation $\phi(z) = \sum_{i=0}^{+\infty} \phi_i \, z^i = \frac{1}{\theta(z)}$ with $\theta(z) = 1 - \sum_{i=0}^{+\infty} \theta^*_i \, z^i$. Equating coefficients of $z_j, j = 0, 1, \dots$, we find that $\phi_0 = 1$ and for $i \ge 1$

$$\phi_i = \sum_{j=1}^{i} \theta^*_j \, \phi_{i-j}.$$

This fact allows us to deduce that the sequences $(\phi_i)_{i\in\mathbb{N}}$ and $(\theta^*_i)_{i\in\mathbb{N}}$ decay at the same rate. Therefore, since $|\theta^*_t| = O\big((t+1)^{-\gamma}\big)$, there exists $h_0 \in \mathbb{Z}$ such that for any $h \ge h_0$, it holds $|\phi_t| \le C\,(t+1)^{-\gamma}$ for some constant $C > 0$. Thus,

$$
\begin{aligned}
r(h) &= \sum_{j=0}^{\infty} \phi_j \, \phi_{j+h} \\
&\leq C^2 \sum_{j=0}^{\infty} \frac{1}{(j+1)^{\gamma}} \, \frac{1}{(j+h+1)^{\gamma}} \\
&\leq C^2 \, (h+1)^{-\gamma} \sum_{j=0}^{\infty} \frac{1}{(j+1)^{\gamma}} \quad \leq C^2 \frac{\pi^2}{6} \, (h+1)^{-\gamma},
\end{aligned}
$$

where the last inequality follows from the fact that $\gamma \geq 2$ and that established the Lemma. □

Lemma 3 *Under assumptions* **A2**, *it holds for any model* $m \in \mathcal{M}_n$, *and for all* $u > 0$

$$
\mathbb{P}\Big(\big\| \widehat{\Sigma}_m^* - \Sigma_m \big\|_{\text{op}} \geq u \Big) \leq 2 \, \exp \left\{ - \frac{s_n}{2} \, \min \left\{ \left(\frac{u}{16 \, D_m \, \sigma_0^2} \right)^2, \frac{u}{32 \, D_m \, \sigma_0^2} \right\} \right\}
$$

Proof One can write for a matrix A

$$
\big\| A \big\|_{\text{op}} = \max_{v \colon \|v\| = 1} \big| v^{\top} A \, v \big| = \big| v_0^{\top} A \, v_0 \big|.
$$

Therefore, one can find a vector $v_0 \in \mathbb{R}^{D_m}$ with $\|v_0\| = 1$ such that

$$
\mathbb{P}\Big(\big\| \widehat{\Sigma}_m^* - \Sigma_m \big\|_{\text{op}} \geq u \Big) = \mathbb{P}\Big(\big| v_0^{\top} \, \big(\widehat{\Sigma}_m^* - \Sigma_m \big) \, v_0 \big| \geq u \Big).
$$

But,

$$
\begin{aligned}
v_0^{\top} \big(\widehat{\Sigma}_m^* - \Sigma_m \big) v_0 &= \frac{1}{n} \sum_{t=1}^{n} \big(v_0^{\top} \, \widehat{\Sigma}_{m,t}^* v_0 - v_0^{\top} \Sigma_m v_0 \big) \\
&= \frac{1}{n} \sum_{t=1}^{n} \big(v_0^{\top} \, (Z_t^{*m}) \, (Z_t^{*m})^{\top} v_0 - v_0^{\top} \Sigma_m v_0 \big) \\
&= \frac{1}{n} \sum_{t=1}^{n} \big(Y_t^2 - \mathbb{E}[Y_t^2] \big)
\end{aligned}
$$

with $Y_t = v_0^{\top} \, Z_t^m = \sum_{i=1}^{D_m} v_0^i \, X_{t-i}^*$. From **A2**, Y_t is $\text{SG}(D_m \, \sigma_0^2)$. Therefore, Y_t^2 is $\text{SE}(256 \, D_m^2 \, \sigma_0^4, 16 \, D_m \sigma_0^2)$ (where SE stands for Sub-Gaussian and SE for Sub-Exponential).

Moreover, we can write

$$
\begin{aligned}
v_0^\top \big(\widehat{\Sigma}_m^* - \Sigma_m\big) v_0 &= \frac{1}{n} \sum_{t=1}^{n} \big(Y_t^2 - \mathbb{E}[Y_t^2]\big) \\
&= \frac{1}{s_n} \sum_{k=0}^{s_n-1} \left(\frac{1}{2q_n} \sum_{i=1}^{q_n} \big(Y_{2kq_n+i}^2 - \mathbb{E}[Y_1^2]\big) \right) \\
&\quad + \frac{1}{s_n} \sum_{k=0}^{s_n-1} \left(\frac{1}{2q_n} \sum_{i=1}^{q_n} \big(Y_{(2k+1)q_n+i}^2 - \mathbb{E}[Y_1^2]\big) \right) \\
&= \mathbf{Y}_1 + \mathbf{Y}_2.
\end{aligned}
$$

Therefore,

$$
\mathbf{Y}_1 = \frac{1}{s_n} \sum_{k=0}^{s_n-1} \mathbf{Y}_{1,k} \quad \text{and} \quad \mathbf{Y}_2 = \frac{1}{s_n} \sum_{k=0}^{s_n-1} \mathbf{Y}_{2,k} \quad \text{with}
$$

$$
\mathbf{Y}_{1,k} = \frac{1}{2q_n} \sum_{i=1}^{q_n} \big(Y_{2kq_n+i}^2 - \mathbb{E}[Y_1^2]\big) \quad \text{and} \quad \mathbf{Y}_{2,k} = \frac{1}{2q_n} \sum_{i=1}^{q_n} \big(Y_{(2k+1)q_n+i}^2 - \mathbb{E}[Y_1^2]\big).
$$

$\{\mathbf{Y}_{1,k}\}$ and $\{\mathbf{Y}_{2,k}\}$ are independent random vectors by virtue of Proposition 4. Now, let us show that $\mathbf{Y}_{i,k}$ are sub-exponentials. For λ such that $|\lambda| < \frac{1}{16\, D_m \sigma_0^2}$, and denoting $w_i = Y_{2kq_n+i}^2 - \mathbb{E}[Y_1^2]$, we have

$$
\begin{aligned}
\mathbb{E}\big[e^{\lambda \mathbf{Y}_{1,k}}\big] &= \mathbb{E}\left[\exp\left(\frac{1}{2q_n} \sum_{i=1}^{q_n} \lambda\, w_i \right) \right] \\
&= \mathbb{E}\left[\Pi_{i=1}^{q_n} \exp\left(\frac{\lambda\, w_i}{2q_n} \right) \right] \\
&= \mathbb{E}\left[\Pi_{i=1}^{q_n} \left(\exp\left(\frac{\lambda\, w_i}{2} \right) \right)^{1/q_n} \right] \\
&\le \Pi_{i=1}^{q_n} \left(\mathbb{E}\left[\exp\left(\frac{\lambda\, w_i}{2} \right) \right] \right)^{1/q_n} \\
&\le e^{\frac{\lambda^2}{2}\, 64\, D_m^2\, \sigma_0^4},
\end{aligned}
$$

where we have used Hölder's Inequality. Hence, $\mathbf{Y}_{1,k}$ is SE($64\, D_m^2\, \sigma_0^4$,$16\, D_m \sigma_0^2$). As a result, using exponential inequalities for SE random variables, it follows

$$
\mathbb{P}\Big(\mathbf{Y}_1 \ge u/2 \Big) \le \exp\left\{ -\frac{s_n}{2}\, \min\left\{ \left(\frac{u}{16\, D_m\, \sigma_0^2} \right)^2, \frac{u}{32\, D_m\, \sigma_0^2} \right\} \right\}
$$

so that

$$
\mathbb{P}\Big(\big| v_0^\top \big(\widehat{\Sigma}_m^* - \Sigma_m\big)\, v_0 \big| \ge u/2 \Big) \le 2\, \exp\left\{ -\frac{s_n}{2}\, \min\left\{ \left(\frac{u}{16\, D_m\, \sigma_0^2} \right)^2, \frac{u}{32\, D_m\, \sigma_0^2} \right\} \right\}.
$$

□

Lemma 4 *Assume* **A3** *holds, then* $\widehat{\Sigma}_m$ *is a.e. invertible. Also,* Σ_m *is invertible.*

Proof We can write $\widehat{\Sigma}_m = \mathbf{M}_m^\top \mathbf{M}_m$ with $\mathbf{M}_m = \left[X_{i-1}, \ldots, X_{i-D_m}\right]_{i=1}^n$. By virtue of **A3**, $\mathbf{M}_m$ is of full rank which implies the a.e. invertibility of $\widehat{\Sigma}_m$.

Moreover, $\Sigma_m = \mathbb{E}\left[\widehat{\Sigma}_m\right] = \mathbb{E}\left[Z_0^m (Z_0^m)^\top\right]$ with $Z_0^m = (X_{-1}, \ldots, X_{-D_m})^\top$. Let $\mathbf{u} \in \mathbb{R}^{D_m}$, it follows $\mathbf{u}^\top \Sigma_m \mathbf{u} = \mathbb{E}\left[((Z_0^m)^\top \mathbf{u})^2\right] \geq 0$. Let show that whenever the equality holds $(\mathbf{u}^\top \Sigma_m = 0)$, $\mathbf{u} = 0$.

Since $((Z_0^m)^\top \mathbf{u})^2 \geq 0$, its expectation vanishes if and only if $(Z_0^m)^\top \mathbf{u} = 0$ a.e. which yields to $\mathbf{u} = 0$ by **A3**. Hence, Σ_m is positive definite and then invertible. □

5 Theoretical tools

The following Proposition that is a consequence of Theorem 3.1 in Doukhan and Wintenberger (2008) gives a link between the τ-mixing coefficients of the process $(X_t)_{t\in\mathbb{Z}}$ and the coefficients θ_i^* of model (3).

Proposition 3 *Assume* **A1** *holds and if* $|\theta_t^*| = O(t^{-\gamma})$ *with* $\gamma > 1$, *there exists a* τ *-weakly dependent stationary solution of* (1) *and a constant* $C_\tau > 0$ *such that for* $r > 0$

$$\tau_{X,\infty}^{(2)}(r) \leq C_\tau \left(\frac{\log r}{r}\right)^{\gamma-1}. \tag{31}$$

Proof With $G(x, \xi_0) = \sigma\, \xi_0 + f_{\theta^*}(x)$ for any $x \in \mathbb{R}^\infty$, it holds

$$\left\|G(x, \xi_0) - G(y, \xi_0)\right\|_2 = \left|f_{\theta^*}(x) - f_{\theta^*}(y)\right| \leq \sum_{i=1}^{\infty} |\theta_i^*|\, |x_i - y_i|.$$

Therefore, (31) is a straightforward application of Theorem 3.1 in Doukhan and Wintenberger (2008). □

The following proposition allows us to obtain the block independence property.

Proposition 4 *Let* $(X_t)_{t\in\mathbb{Z}}$ *be the stationary mixing process obtained in Proposition 3. Let also* s_n, q_n, A_k, B_k *defined as above for* $k = 0, \ldots, s_n - 1$. *There exist random vectors* $A_k^* = \left(\mathbf{X}_{2kq_n+1}^*, \ldots, \mathbf{X}_{(2k+1)q_n}^*\right)$, $B_k^* = \left(\mathbf{X}_{(2k+1)q_n+1}^*, \ldots, \mathbf{X}_{(2k+2)q_n}^*\right)$ *such that*:

1. *For* $k = 0, \ldots, s_n - 1$, A_k^* *has the same law as* A_k, *also* B_k^* *and* B_k.
2. *The random vectors* $(A_k^*)_{0\leq k\leq s_n-1}$ *are independent and so are the vectors* $(B_k^*)_{0\leq k\leq s_n-1}$.

3.
$$\left\|A_k - A_k^*\right\|_1 \leq q_n K_n \tau_{X,\infty}^{(1)}(q_n)$$
$$\text{and } \left\|B_k - B_k^*\right\|_1 \leq q_n K_n \tau_{X,\infty}^{(1)}(q_n).$$

The next Theorem is a Talagrand's Inequality given in Klein and Rio (2005).

Theorem 2 *Let $Y_1, \ldots, Y_n$ be independent random variables and let $\mathcal{F}$ be a countable class of uniformly bounded measurable functions. Then, for all $\alpha > 0$,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} |\eta_n(g)|^2 - 2(1+2\alpha)H^2\right]_+ \leq \frac{2}{K}\left(\frac{v}{n}e^{-K\alpha\frac{nH^2}{v}} + \frac{49M^2}{4Kn^2C^2(\alpha)}e^{-\frac{2\sqrt{2}KC(\alpha)\sqrt{\alpha}}{7\sqrt{2}}\frac{nH}{M}}\right)$$

with $\eta_n(g) = n^{-1}\sum_{t=1}^n (g(Y_t) - \mathbb{E}[g(Y_t)])$ for any $g \in \mathcal{F}$;

$$C(\alpha) = (\sqrt{1+\alpha} - 1) \wedge 1, K = 1/6$$

$$\sup_{g\in\mathcal{F}} \|g\|_\infty \leq M, \quad \mathbb{E}\left[\sup_{g\in\mathcal{F}} |\eta_n(g)|\right] \leq H, \quad \sup_{g\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n \text{Var}\,(g(Y_t)) \leq v.$$

Acknowledgements The author thanks William KENGNE and Jean-Marc BARDET for proofreads and helpful discussions. Kare Kamila has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754362.

References

Baraud, Y., Comte, F., Viennet, G. (2001a). Model selection for (auto-) regression with dependent data. *ESAIM*: *Probability and Statistics, 5,* 33–49.

Baraud, Y., Comte, F., Viennet, G., et al. (2001b). Adaptive estimation in autoregression or-mixing regression via model selection. *The Annals of Statistics, 29*(3), 839–875.

Bardet, J.-M., Wintenberger, O. (2009). Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. *The Annals of Statistics, 37*(5B), 2730–2759.

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

Birgé, L., Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields, 138*(1–2), 33–73.

Comte, F., Dedecker, J., Taupin, M.-L. (2008). Adaptive density deconvolution with dependent inputs. *Mathematical methods of Statistics, 17*(2), 87.

Comte, F., Genon-Catalot, V. (2020). Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics, 72*(4), 1023–1054.

Dedecker, J., Prieur, C. (2005). New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields, 132*(2), 203–236.

Doukhan, P., Wintenberger, O. (2008). Weakly dependent chains with infinite memory. *Stochastic Processes and their Applications, 118*(11), 1997–2013.

Goldenshluger, A., Zeevi, A. (2001). Nonasymptotic bounds for autoregressive time series modeling. *Annals of Statistics, 29,* 417–444.

Hsu, D., Kakade, S. M., Zhang, T. (2011). An analysis of random design linear regression. arXiv preprint arXiv:1106.2363 .

Ing, C.-K., Wei, C.-Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis, 85*(1), 130–155.

Ing, C.-K., Wei, C.-Z., et al. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics, 33*(5), 2423–2474.

Klein, T., Rio, E., et al. (2005). Concentration around the mean for maxima of empirical processes. *The Annals of Probability, 33*(3), 1060–1077.

Lebarbier, E., Mary-Huard, T. (2004). Le critère BIC: fondements théoriques et interprétation, PhD thesis, INRIA.

Lerasle, M., et al. (2011). Optimal model selection for density estimation of stationary data under various mixing conditions. *The Annals of Statistics, 39*(4), 1852–1877.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics, 8*(1), 147–164.

Van de Geer, S. A. (2002). On hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data* (pp. 161–169), Springer.