DATA DRIVEN MODEL SELECTION FOR SAME-REALIZATION PREDICTIONS IN AUTOREGRESSIVE PROCESSES

Kare Kamila

DATA DRIVEN MODEL SELECTION FOR SAME-REALIZATION PREDICTIONS IN AUTOREGRESSIVE PROCESSES

BY Kare KAMILA*

March 15, 2021

SAMM, Université Paris 1 Panthéon-Sorbonne, FRANCE

Abstract

This paper is about the one-step ahead prediction of the future of observations drawn from an infinite-order autoregressive AR(∞) process. The aim of this paper is to design penalties (complete data driven) ensuring that the selected model verifies the efficiency property but in the non asymptotic framework. We present an oracle inequality with a leading constant equal to one. Moreover, we also 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 $1 - c_0/n^3$ 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. Monte Carlo experiments are performed to highlight the obtained results.

Key words: Model selection, oracle inequality, efficiency, autoregressive process, data driven.

1 Introduction

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

$$X_t = f^*\big((X_{t-i})_{i\in\mathbb{N}^*}\big) + \sigma\,\xi_t \quad \text{for any } t \in \mathbb{Z}. \tag{1.1}$$

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.

The problem is to estimate the function f^* using these observations. The process (1.1) is a particular case of the general class of affine causal process studied in [10] and [4]. 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:

$$\sum_{k=1}^{\infty}\left(\sup_{x\in\mathbb{R}^\infty}\Big|\frac{\partial}{\partial x_k}f^*(x)\Big|\right) < 1, \tag{1.2}$$

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

provided that that f^* admits partial derivatives on $\mathbb{R}^{\mathbb{N}}$. Under (1.2) and if the noise ξ_0 admits r-order moments, [10] showed that there exists a stationary, mixing and ergodic solution to (1.1) admitting r-order moments.
Moreover, [4] studied the consistency and the asymptotic normality of the 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

$$f^*(X_{t-1}, X_{t-2}, \ldots) = f_{\theta^*}(X_{t-1}, X_{t-2}, \ldots) = \sum_{i=1}^{\infty} \theta_i^* X_{t-i}. \tag{1.3}$$

For such processes, condition (1.2) becomes

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

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 almost all invertible causal ARMA(p, q) processes and it is very useful for prediction given the past. Moreover, contrary to the autocariance of ARMA(p, q) processes which decays exponentially fast, AR(∞) are able to model more complex behaviour 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 (1.1) 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 $t < 0$. However, θ^* is generally unknown and it is impossible to provide a direct estimator since its coordinate 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. [17] 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, [13] and [15] 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. [11] in the non parametric framework, studied how well a Gaussian process admitting an AR(∞) representation can be approximated by a finite-order AR model.

In [2] and [3], 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 [7]) and be bounded by a function of the model parameter θ^* (see [10]). In this work, we do not assume any mixing property of the process since the condition **A1** implies the τ-mixing property (see [10]) 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 mimic the oracle when observations arise form AR(∞) under mild conditions, including the existence of the all order moment of the noise, the decreasing rate of the coefficients of $(\theta_i^*)_{i\in\mathbb{N}}$ so that thanks to a result by [10], the generating process has nice properties such as stationarity, τ-mixing.
The main contributions of this paper include:
(i) Using least squares contrast, we have shown an oracle inequality with a leading constant equal to one.
(ii) We have also proved 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 are described in Section 2. The main results are presented Section 3. Finally, numerical results are presented in Section 4 and Section 5 contains the proofs.

2 Model Selection Approach and Preliminary Results

2.1 Model Selection Approach

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

$$f(x_1, x_2, \ldots, x_{D_m}) = \sum_{i=1}^{D_m} \theta_i\, x_i, \tag{2.1}$$

with $\theta = (\theta_1, \ldots, \theta_{D_m}) \in \Theta_m$ and Θ_m a compact set of $\mathbb{R}^{D_m}$. 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

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

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

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

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

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

Its empirical version minimizing the least-squares contrast is

$$\widehat{\theta}_m = \operatorname*{argmin}_{\theta\in\Theta_m} \gamma_n(\theta) \quad \text{where} \quad \gamma_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}(X_t - f_\theta^t)^2. \tag{2.2}$$

Note that (thanks to stationarity of the process), for an estimate $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$, the excess loss can be rewritten as

$$\ell(\widehat{\theta}, \theta^*) = \mathbb{E}\Big[\big\|F_{\widehat{\theta}} - F_{\theta^*}\big\|_n^2\Big] \tag{2.3}$$

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

$$\widehat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \{C(m)\} \qquad \text{with} \qquad C(m) := \gamma_n\big(\widehat{\theta}_m\big) + \operatorname{pen}(S_m). \tag{2.4}$$

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

$$m^* \in \arg \inf_{m \in \mathcal{M}_n} \ell(\widehat{\theta}_m, \theta^*). \tag{2.5}$$

The oracle m^* is unachievable since it depends on θ^* and the distribution $P_{(X_1,\dots,X_n)}$ 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 twofold. First, we want to propose a data driven penalty in order to obtain an oracle inequality

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1 \inf_{m \in \mathcal{M}_n} \big\{\ell(\widehat{\theta}_m, \theta^*)\big\} + \frac{C_2}{n} \tag{2.6}$$

with the leading constant C_1 close to one and $C_2 > 0$.
Since for every $m \in \mathcal{M}_n$, we have the following decomposition holds

$$\begin{aligned} \ell(\widehat{\theta}_m, \theta^*) &= \ell(\theta^*_m, \theta^*) + \ell(\widehat{\theta}_m, \theta^*_m) \\ &=: \text{Biais}(m) + \text{Variance}(m), \end{aligned} \tag{2.7}$$

the inequality (2.6) implies 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. 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.

Moreover, in decomposition (2.7), the bias term is generally not irreducible but the variance can be approximately proportional to the size of the model. Our second goal is to simplify the oracle inequality (2.6) in order to obtain the following common inequality

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1' \inf_{m \in \mathcal{M}_n} \left\{\ell(\theta^*_m, \theta^*) + \operatorname{pen}(S_m)\right\} + \frac{C_2'}{n} \tag{2.8}$$

with the leading constant $C_1' = 1 + \delta$ with $\delta > 0$ (and close to 0) and $C_2' > 0$.

2.2 Notations

We will use the following norms:

- $\|.\|$ denotes the usual Euclidean norm on $\mathbb{R}^\nu$, with $\nu \geq 1$;
- $\|A\|_{\text{op}}$ 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 $\|X\|_r = \big(\mathbb{E}\big[\|X\|^r\big]\big)^{1/r} \in [0, \infty]$.

2.3 Preliminary Results

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 [9]. 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 $|f(x) - f(y)| \le \|x - y\|_E$.

Using the 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 $s \ge 0$, let introduce the norm $\|x - y\|_{\mathbb{R}^k} = (|x_1 - y_1| + \cdots + |x_k - y_k|)$ and setting $\mathcal{M}_i = \sigma(Z_t, t \le i)$ and if $\mathbb{E}(|Z_1|) < \infty$, let

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

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.

The next Proposition that is a consequence of Theorem 3.1 in [10] gives a link between the τ-mixing coefficients of the process $(X_t)_{t\in\mathbb{Z}}$ and the coefficients θ^*_i of the model (1.3).

Proposition 1. *Assume **A1** holds and if $|\theta^*_t| = O(t^{-\gamma})$ with $\gamma > 1$, there exists a τ-weakly dependent stationary solution of (1.1) and a constant $C_\tau > 0$ such that for $r > 0$*

$$\tau^{(2)}_{X,\infty}(r) \le C_\tau \Big(\frac{\log r}{r}\Big)^{\gamma-1} \tag{2.9}$$

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

$$\big\|G(x, \xi_0) - G(y, \xi_0)\big\|_2 = \big|f_{\theta^*}(x) - f_{\theta^*}(y)\big| \le \sum_{i=1}^{\infty} |\theta^*_i|\,|x_i - y_i|.$$

Therefore (2.9) is a straightforward application of Theorem 3.1 in [10]. ∎

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

Now, we construct random variables approximating $\vec{X}_t$'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

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

We recall a result of [16] which is a consequence of the coupling in [9].

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

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

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.

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

Condition **A2** implies that the vector $Z_t^m = (X_{t-1}, \dots, X_{t-D_m})^\top$ 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 $\|v\| = 1$,

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(\lambda\, v^\top Z_t^m\big)\Big] &= \mathbb{E}\left[\prod_{i=1}^{D_m} \exp\Big(\lambda\, v_i(X_{t-i}\Big)\right] \\
&\le \prod_{i=1}^{D_m} \Big\| \exp\Big(\lambda\, v_i(X_{t-i}\Big)\Big\|_{D_m} \\
&= \prod_{i=1}^{D_m} \exp\big(\lambda^2\, D_m\, \sigma_0^2\, v_i^2/2\big) \\
&= e^{\frac{\lambda^2}{2}\, D_m\, \sigma_0^2},
\end{aligned}$$

where the Inequality follows from Hölder's Inequality.

The following assumption provides a sufficient condition to ensure the invertibility of both $\widehat{\Sigma}_m$ and Σ_m.

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 $\gamma \ge 1$. 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 $\inf\limits_{-\pi\le\lambda<\pi} g(\lambda) \ge a$.

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

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

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

$$\begin{aligned} g(\lambda) &= \frac{\sigma^2}{2\pi\left|1-\theta_1^*\,e^{-i\lambda}\right|^2} \\ &= \frac{\sigma^2}{2\pi\Big(1-2\,\theta_1^*\,\cos(\lambda)+(\theta_1^*)^2\Big)}, \end{aligned}$$

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 $-1 \leq \cos(x) \leq 1$ for any real x, it follows for every λ

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

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

$$\big\|\Sigma_m\big\|_{\rm op} \leq \pi^{-1}\sum_{i=0}^{\infty}\big|\mathbb{E}[X_0\,X_i]\big| < \infty. \tag{2.10}$$

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

$$\big\|\Sigma_m^{-1}\big\|_{\rm op} \leq 1/a. \tag{2.11}$$

Now for technical convenience, we choose the integer s_n and the cardinal of $\mathcal{M}_n$ such that

$$\mathbf{A5}: \quad \frac{a\,s_n}{2}\,\min\left\{\left(\frac{(r\wedge 1)}{2^6\,\sigma_0^2\,K_n}\right)^2, \frac{(r\wedge 1)}{2^7\,\sigma_0^2\,K_n}\right\} \geq 3\log n, \tag{2.12}$$

where $r := a/\mathbb{E}[X_0^2]$ and $b\wedge c$ is the minimum of b and c. This means that s_n is of the form $s_n = C\,\log n$ where $C \geq 6\,a^{-1}\max\left\{\left(\frac{2^6\,\sigma_0^2\,K_n}{(r\wedge 1)}\right)^2, \frac{2^7\,\sigma_0^2\,K_n}{(r\wedge 1)}\right\}$. For instance, if $K_n = \log n$

and $a \approx \mathbb{E}[X_0^2] \approx \sigma_0^2$, we can choose $s_n = 6\,(2)^{12}\,(\log)^3 n$. Henceforth, in all the sequel K_n will satisfy

$$\mathbf{A6}: \quad K_n = C_K \log n \quad \text{for some constant } C_K > 0. \tag{2.13}$$

Let us introduce extra important notations. From the definition of the LSE (2.2), it follows that

$$\widehat{\theta}_m = \widehat{\Sigma}_m^{-1} \mathbf{M}_m^\top X \tag{2.14}$$

where the matrix $\mathbf{M}_m = \big[X_{i-1}, \ldots, X_{i-D_m}\big]_{i=1}^n$, $\widehat{\Sigma}_m = \mathbf{M}_m^\top \mathbf{M}_m$ and $X = (X_1, \ldots, X_n)^\top$, provided that $\widehat{\Sigma}_m$ is invertible almost everywhere (see Lemma 6). Let denote the expected value of the random matrix $\widehat{\Sigma}_m$ by $\Sigma_m = \mathbb{E}\big[\widehat{\Sigma}_m\big]$. Rewriting (1.1) in a vectorial form (with $\xi = (\xi_1, \ldots, \xi_n)^\top$), i.e. $X = F_{\theta^*} + \sigma\,\xi = (F_{\theta^*} - F_{\theta_m^*}) + F_{\theta_m^*} + \sigma\,\xi$, it follows that

$$\mathbf{M}_m^\top X = \mathbf{M}_m^\top (F_{\theta^*} - F_{\theta_m^*}) + \mathbf{M}_m^\top \mathbf{M}_m \theta_m^* + \sigma\,\mathbf{M}_m^\top \xi.$$

Thus, using (2.14) it holds

$$\widehat{\theta}_m - \theta_m^* = \widehat{\Sigma}_m^{-1} \mathbf{M}_m^\top (F_{\theta^*} - F_{\theta_m^*}) + \sigma\,\widehat{\Sigma}_m^{-1} \mathbf{M}_m^\top \xi, \tag{2.15}$$

which implies

$$\begin{aligned} \mathbf{M}_m(\widehat{\theta}_m - \theta_m^*) &= \mathbf{M}_m \widehat{\Sigma}_m^{-1} \mathbf{M}_m^\top (F_{\theta^*} - F_{\theta_m^*}) + \sigma\,\mathbf{M}_m \widehat{\Sigma}_m^{-1} \mathbf{M}_m^\top \xi \\ &= \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta_m^*}\big) + \sigma\,\mathbf{P}_{\mathbf{M}_m}(\xi) \end{aligned}$$

where $\mathbf{P}_{\mathbf{M}_m} = \mathbf{M}_m\big(\mathbf{M}_m^\top \mathbf{M}_m\big)^{-1} \mathbf{M}_m^\top$ is the projection matrix onto the sub-space spanned by the columns of $\mathbf{M}_m$.
The main difficulty in this work lies in the handling of the matrix $\widehat{\Sigma}_m^{-1}$. We are going to use a classical approach to overcome this issue, it consists in defining a set on which the $\widehat{\Sigma}_m^{-1}$ can be approximated by Σ_m^{-1} ([3],[12], [19],[8] among other) which is invertible (see Lemma 6). For $m \in \mathcal{M}_n$, let define $\Gamma_{m,r}$ the set

$$\Gamma_{m,r} = \Big\{ \big\|\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1}\big\|_{\text{op}} \le r\,\big\|\Sigma_m^{-1}\big\|_{\text{op}} \Big\}.$$

We will see that in our framework $\Gamma_{m,r}$ holds with high probability. Before proving that, let us notice that $\widehat{\Sigma}_m$ can be rewritten as

$$\widehat{\Sigma}_m = \frac{1}{n}\sum_{t=1}^n \widehat{\Sigma}_{m,t} \quad \text{with} \quad \widehat{\Sigma}_{m,t} = Z_t^m (Z_t^m)^\top$$
$$\text{where} \quad Z_t^m = (X_{t-1}, \ldots, X_{t-D_m})^\top. \tag{2.16}$$

Since we have fixed r equal to $a/\mathbb{E}[X_0^2]$, the set $\Gamma_{m,r}$ will be denoted by Γ_m in the all the sequel.

The following result shows that the event Γ_m holds with high probability.

Proposition 4. *Under assumptions **A1** – **A6** and if $|\theta_t^*| = O(t^{-\gamma})$ with $\gamma \ge 8$, it holds*

$$\mathbb{P}(\Gamma_m^c) \le \frac{c_0}{n^3}, \tag{2.17}$$

with $c_0 = 1 + 8\,A\,C_\tau\,C_K\,\big\|X_0\big\|_2\,(a\,(r \wedge 1))^{-1}$ where A satisfies (5.1).

3 Oracle Inequality for Same Realization Prediction

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

Theorem 3.1. *Let consider observations $(X_1, \ldots, X_n)$ arising from a solution of the process (1.1) 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 $x \geq 4$, let a penalty function pen: $\mathcal{M}_n \to \mathbb{R}^+$ such that*

$$\text{pen}(S_m) = x\,\sigma^2\,\frac{D_m}{n}. \tag{3.1}$$

Then with probability at least $1 - c_0\,n^{-2}$, the LSE $\widehat{\theta}_{\widehat{m}}$ with $\widehat{m}$ given in (2.4), satisfies

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq \inf_{m \in \mathcal{M}_n} \left\{ \ell(\widehat{\theta}_m, \theta^*) \right\} + \frac{2\,x\,\sigma^2}{n}. \tag{3.2}$$

Let us give some remarks about this result:

- The oracle inequality (3.2) is optimal in the sense that the leading constant is exactly one
- This result is new in non asymptotic framework for AR(∞) under mild conditions. Indeed, [14] obtained a counterpart of our result when $n \to \infty$ under several assumptions;
- The designed penalty (3.1) generalizes the Mallows C_p.

We can obtain for free as a consequence of Theorem 3.1, the asymptotic efficiency obtained by [18] and [15].

Corollary 1. *Under the assumptions of Theorem 3.1, it holds*

$$\frac{\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)}{\inf\limits_{m \in \mathcal{M}_n} \left\{ \ell(\widehat{\theta}_m, \theta^*) \right\}} \xrightarrow[n \to \infty]{\mathcal{P}} 1.$$

At present, we state the second main result describe in (2.8).

Theorem 3.2. *Under the assumptions of the Theorem 3.1, with the same penalty (3.1), then with probability at least $1 - c_0\,n^{-2}$, the LSE $\widehat{\theta}_{\widehat{m}}$ with $\widehat{m}$ given in (2.4), satisfies*

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq 2 \inf_{m \in \mathcal{M}_n} \left\{ \ell(\theta_m^*, \theta^*) + \text{pen}(S_m) \right\} + \frac{2\,x\,\sigma^2}{n}. \tag{3.3}$$

Let us comment the optimality of the leading constant 2 relatively to the ones obtained in similar framework :

1. for regression in fixed design, Birgé and Massart ([5], [6]) obtained a similar inequality with the leading constant $C = 1 + \delta$ with $\delta > 0$. Also in [1], $C = 1 + \delta$ with $\delta \in (0, 2)$.
2. in Theorem 3.1 in [2], the leading constant is equal to 4 and worth $2.\frac{(x+\rho)^2}{(x-\rho)^2} > 2$ in [3]. These results have been obtained in a more general framework but with strong conditions.

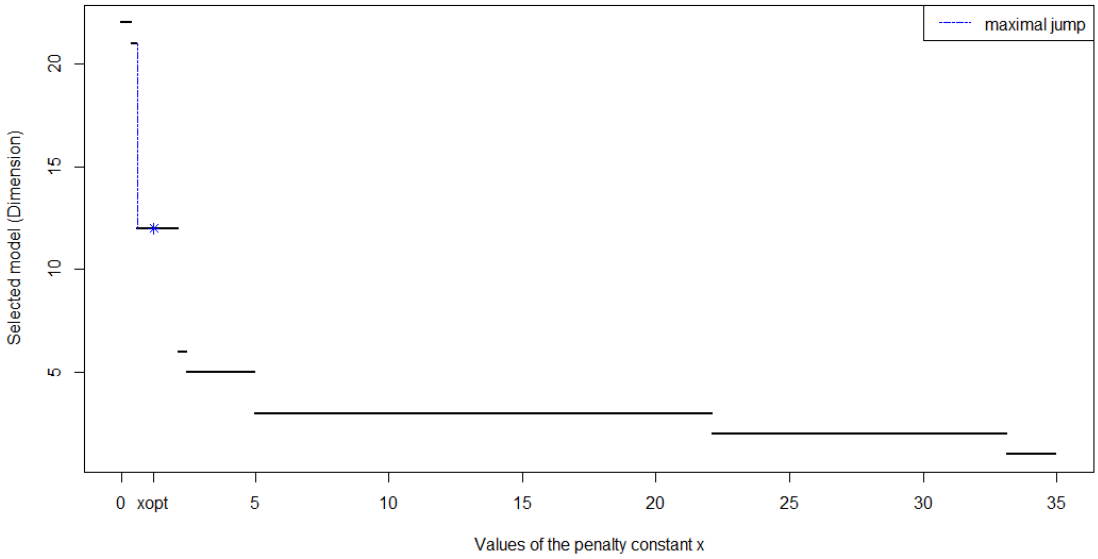

Figure 1: Dimension Jump

4 Numerical Experiments

This section aims at investigating how well the found penalties are in accordance with the results in Section 3. To do that, we generated observations from a causal invertible ARMA$(1,1)$

$$X_t = \phi_0\, X_{t-1} + \xi_t + \theta_0\, \xi_{t-1}$$

where the ξ_t' s are independent and identically $\mathcal{N}(0,1)$ distributed and

$$(\phi_0, \theta_0) \in \Big\{(a,b) : a \in \{0.9, 0.7, 0.5, -0.9, -0.7, -0.5\} \quad \text{and} \quad b \in \{0.8, 0.6, -0.8, -0.6\}\Big\}$$

as in [15].

Since, all of these models are invertibles, then they admit an AR(∞) representation. In order to attest our theoretical results, we consider as candidate set of models, the family $\mathcal{M}_n$ of increasing AR(p) defined in (2.1) with $1 \le p \le K_n$ where $K_n = \lfloor 2 \log n \rfloor$ according to condition **A6**.

For each pair (ϕ_0, θ_0), we compute an empirical version of

$$ME := \frac{\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)}{\inf\limits_{m \in \mathcal{M}_n} \{\ell(\widehat{\theta}_m, \theta^*)\}}$$

with $\widehat{m}$ selected as in (2.4) where $\text{pen}(S_m) = \widehat{x}\,\sigma^2\,\frac{D_m}{n}$ and the optimal constant $\widehat{x}$ has been calibrated using the dimension jump algorithm implemented in R capushe package and illustrated in Figure 1. In order to do so and produce Figure 1, we used an ARMA(1,1) model with $\phi_0 = 0.9$ and $\theta_0 = 0.7$ and a sample size of 500. Then, $\widehat{x} = 2\,x$ where x is the value which gives the highest jump. In Figure 1, $\widehat{x} = \text{xopt} = 2 * 0.5975 = 1.195$. This optimal value was set throughout the simulation study.

In the penalty $\text{pen}(S_m)$, the variance σ^2 is estimated by considering the largest model, i.e. of size K_n as traditionally done with Mallows' Cp. The Table 1 summarizes the obtained results over 500 replications.

Table 1: Empirical estimates of ME

		θ_0			
ϕ_0	n/K_n	**0.8**	**0.6**	**-0.8**	**-0.6**
0.9	60/8	1.17	1.24	1.21	1.11
	120/9	1.10	1.19	1.12	1.06
	200/10	1.05	1.16	1.11	1.12
	500/12	1.01	1.04	1.13	1.05
	1000/13	1.01	1.01	1.04	1.06
	2000/15	1.00	1.01	1.03	1.03
0.7	60/8	1.20	1.23	1.15	1.22
	120/9	1.15	1.21	1.09	1.16
	200/10	1.14	1.18	1.18	1.24
	500/12	1.03	1.15	1.09	1.16
	1000/13	1.01	1.07	1.11	1.12
	2000/15	1.01	1.03	1.03	1.12
0.5	60/8	1.25	1.13	1.12	1.14
	120/9	1.13	1.11	1.06	1.14
	200/10	1.15	1.14	1.07	1.16
	500/12	1.08	1.14	1.05	1.10
	1000/13	1.03	1.08	1.03	1.10
	2000/15	1.01	1.07	1.04	1.11
-0.9	60/8	1.14	1.09	1.17	1.21
	120/9	1.10	1.10	1.07	1.14
	200/10	1.10	1.11	1.04	1.14
	500/12	1.10	1.12	1.02	1.04
	1000/13	1.04	1.08	1.01	1.02
	2000/15	1.07	1.04	1.00	1.01
-0.7	60/7	1.17	1.21	1.20	1.17
	120/8	1.14	1.21	1.11	1.11
	200/9	1.18	1.12	1.12	1.21
	500/12	1.11	1.12	1.03	1.13
	1000/13	1.06	1.05	1.01	1.07
	2000/15	1.05	1.11	1.00	1.03
-0.5	60/8	1.10	1.22	1.13	1.12
	120/9	1.12	1.13	1.14	1.12
	200/10	1.09	1.08	1.14	1.08
	500/12	1.06	1.17	1.11	1.08
	1000/13	1.05	1.12	1.04	1.06
	2000/15	1.04	1.07	1.01	1.09

As can be seen, these results confirm our theory because we can observe that $\widehat{ME}$ is not far from 1 and is decreasing with n. Moreover, we note the convergence towards 1 as announced in our theoretical results. These results are better than those obtained in [15].

The fastest decreasing rate occurs when $\phi_0 \in \{0.9, -0.9\}$ and $\theta_0 \in \{0.8, -0.8\}$ with $sgn(\phi_0) = sgn(\theta_0)$ (where $sgn(a) = 1$ if $a > 0$ and $sgn(a) = -1$ otherwise). Let us also note that even if there is a global decrease between $n = 60$ and $n = 2000$, there are cases where $\widehat{ME}$ increases slightly from $n = 1000$. This happens very often when ϕ_0 is very close to $-\theta_0$ and that makes the process $(X_t)_t$ very close to a white noise which is unpredictable. We think that this is what justifies the global non-decay in these cases.

5 Proofs

5.1 Proof of Theorem 3.1

Proof.

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

Since $\ell(\widehat{\theta}_m, \theta^*) = \mathbb{E}\Big[\frac{1}{n}\sum_{t=1}^{n}(f^t_{\theta^*} - f^t_{\widehat{\theta}_m})^2\Big]$ by stationarity, the difficult part of the proof is to optain the penalty term from the expectation of the scalar product $< \xi, (f_{\widehat{\theta}_m} - f_{\theta^*}) >_n$. Let $m \in \mathcal{M}_n$. By definition $C(\widehat{m}) \leq C(m)$. Therefore

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

Moreover, the term of interest can be decomposed into

$$\begin{aligned}< \xi, F_{\widehat{\theta}_m} - F_{\theta^*} > &= < \xi, F_{\widehat{\theta}_m} - F_{\theta^*_m} > + < \xi, F_{\theta^*_m} - F_{\theta^*} > \\ &= \sigma\, < \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) > + < \xi, \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big) > + < \xi, F_{\theta^*_m} - F_{\theta^*} > \\ &= \sigma\, < \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) > + < \xi, (\mathbf{I}_n - \mathbf{P}_{\mathbf{M}_m})\big(F_{\theta^*_m} - F_{\theta^*}\big) > .\end{aligned}$$

and then

$$\begin{aligned}&\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\big\|_n^2 - 2\,\sigma^2 < \xi, \mathbf{P}_{\mathbf{M}_{\widehat{m}}}(\xi) >_n - 2\,\sigma < \xi, (\mathbf{I}_n - \mathbf{P}_{\mathbf{M}_{\widehat{m}}})\big(F_{\theta^*_{\widehat{m}}} - F_{\theta^*}\big) >_n + \mathrm{pen}(S_{\widehat{m}}) \\ \leq\ &\big\|F_{\widehat{\theta}_m} - F_{\theta^*}\big\|_n^2 - 2\,\sigma^2 < \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) >_n - 2\,\sigma < \xi, (\mathbf{I}_n - \mathbf{P}_{\mathbf{M}_m})\big(F_{\theta^*_m} - F_{\theta^*}\big) >_n + \mathrm{pen}(S_m)\end{aligned}$$

Therefore, it is quite easy to obtain the desired expectation. But since the matrix $\mathbf{P}_{\mathbf{M}_m}$ is random and not independent of ξ, the task is difficult. Taking expectation and applying Lemma 5, it yields

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) + \big(\mathrm{pen}(S_{\widehat{m}}) - 2\,\sigma^2\,\mathbb{E}\big[< \xi, \mathbf{P}_{\mathbf{M}_{\widehat{m}}}(\xi) >_n\big]\big) \leq \ell(\widehat{\theta}_m, \theta^*) + \big(\mathrm{pen}(S_m) - 2\,\sigma^2\,\mathbb{E}\big[< \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) >_n\big]\big).$$

In view of Lemma 4, and the choice of the penalty according to (3.1), it holds on Γ_m

$$0 \leq \mathrm{pen}(S_m) - 2\,\sigma^2\,\mathbb{E}\big[< \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) >_n\big] \leq 2\,\mathrm{pen}(S_m).$$

As a result on Γ_m,

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq \ell(\widehat{\theta}_m, \theta^*) + 2\,\mathrm{pen}(S_m).$$

Let set $\Gamma = \bigcap_{m\in\mathcal{M}_n}\Gamma_m$. Γ holds with probability larger than $1 - c_0\,n^{-2}$. Indeed,

$$\mathbb{P}(\Gamma) = 1 - \mathbb{P}\Big(\bigcup_{m\in\mathcal{M}_n}\Gamma_m^c\Big) \geq 1 - \sum_{m\in\mathcal{M}_n}\mathbb{P}(\Gamma_m^c) \geq 1 - \frac{c_0\,K_n}{n^3} \geq 1 - \frac{c_0}{n^2},$$

using Proposition 4 and the fact that $K_n \leq n$. Hence, it holds on Γ

$$\begin{aligned}\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) &\leq \inf_{m\in\mathcal{M}_n}\big\{\ell(\widehat{\theta}_m, \theta^*) + 2\,\mathrm{pen}(S_m)\big\} \\ &\leq \inf_{m\in\mathcal{M}_n}\big\{\ell(\widehat{\theta}_m, \theta^*)\big\} + \frac{2\,x\,\sigma^2}{n}\end{aligned}$$

since the smallest dimension is one. That ends the proof. ∎

5.2 Proof of Theorem 3.2

Proof. We have

$$\begin{aligned}\ell(\widehat{\theta}_m, \theta^*_m) &= \frac{1}{n}\,\mathbb{E}\Big[\sum_{t=1}^{n}(f^t_{\widehat{\theta}_m} - f^t_{\theta^*_m})^2\Big] \\ &= \frac{1}{n}\,\mathbb{E}\big[(\widehat{\theta}_m - \theta^*_m)^\top\,\mathbf{M}_m^\top\mathbf{M}_m\,(\widehat{\theta}_m - \theta^*_m)\big].\end{aligned}$$

Using (2.15), it follows

$$\begin{aligned}(\widehat{\theta}_m - \theta^*_m)^\top\,\mathbf{M}_m^\top\mathbf{M}_m\,(\widehat{\theta}_m - \theta^*_m) =& < \big(F_{\theta^*} - F_{\theta^*_m}\big), \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big) > \\ &+ \sigma^2\ < \xi, \mathbf{P}_{\mathbf{M}_m}\,\xi > + 2\sigma\ < \xi, \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*_m} - F_{\theta^*}\big) > .\end{aligned}$$

Moreover since $\mathbf{P}_{\mathbf{M}_m}$ is a projection matrix $\big\|\mathbf{P}_{\mathbf{M}_m}\big\|_{\text{op}} = 1$ and,

$$\begin{aligned}\mathbb{E}\big[|\ < \big(F_{\theta^*} - F_{\theta^*_m}\big), \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big) > |\big] &\le \mathbb{E}\Big[\big\|F_{\theta^*} - F_{\theta^*_m}\big\|\,\big\|\mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big)\big\|\Big] \\ &\le \mathbb{E}\Big[\big\|F_{\theta^*} - F_{\theta^*_m}\big\|\,\big\|\mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big)\big\|\Big] \\ &\le \mathbb{E}\Big[\big\|F_{\theta^*} - F_{\theta^*_m}\big\|^2\Big]\end{aligned}$$

Hence, we deduce from Lemma 4

$$\begin{aligned}\ell(\widehat{\theta}_m, \theta^*_m) &= \frac{1}{n}\,\mathbb{E}\big[\ < \big(F_{\theta^*} - F_{\theta^*_m}\big), \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*} - F_{\theta^*_m}\big) > \big] + \frac{\sigma^2}{n}\,\mathbb{E}\big[\ < \xi, \mathbf{P}_{\mathbf{M}_m}\,\xi > \big] \\ &\le \ell(\theta^*_m, \theta^*) + \frac{2}{n}\,\sigma^2\,D_m \\ &\le \ell(\theta^*_m, \theta^*) + 0.5\,\text{pen}(S_m).\end{aligned}$$

So that

$$\begin{aligned}\ell(\widehat{\theta}_m, \theta^*) &= \ell(\theta^*_m, \theta^*)\ +\ \ell(\widehat{\theta}_m, \theta^*_m) \\ &\le 2\Big(\ell(\theta^*_m, \theta^*) + \text{pen}(S_m)\Big)\end{aligned}$$

This fact along with Theorem 3.1 establishes (3.3). ■

5.3 Proof of Corollary 1

Proof. First, let remark that $\mathbb{P}(\Gamma) \underset{n\to\infty}{\longrightarrow} 1$ (where Γ is the set defined in the proof of Theorem 3.1). Also, from (3.2) and for any n

$$\inf_{m\in\mathcal{M}_n}\big\{\ell(\widehat{\theta}_m, \theta^*)\big\} \le \ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \le \inf_{m\in\mathcal{M}_n}\big\{\ell(\widehat{\theta}_m, \theta^*)\big\} + \frac{2\,x\sigma^2}{n}.$$

The proof is done after considering $n \to \infty$ in the previous double-inequality. ■

5.4 Proof of Proposition 4

Let recall the definition of $\widehat{\Sigma}_m$ as in (2.16),

$$\widehat{\Sigma}_m = \frac{1}{n}\sum_{t=1}^{n}\widehat{\Sigma}_{m,t} \quad \text{with} \quad \widehat{\Sigma}_{m,t} = Z^m_t(Z^m_t)^\top \text{where} \quad Z^m_t = (X_{t-1}, \ldots, X_{t-D_m})^\top.$$

Following idea proof of Proposition 4 in [8], we claim that

$$\Gamma^c_r = \Big\{\big\|\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1}\big\|_{\text{op}} > r\,\big\|\Sigma_m^{-1}\big\|_{\text{op}}\Big\} \ \subset\ \Big\{\big\|\Sigma_m^{-1/2}\,\widehat{\Sigma}_m\,\Sigma_m^{-1/2} - \mathbf{I}_m\big\|_{\text{op}} > \frac{r\wedge 1}{2}\Big\},$$

so that

$$\Gamma_r^c \subset \Big\{ \big\|\widehat{\Sigma}_m - \Sigma_m\big\|_{\text{op}} \big\|\Sigma_m^{-1}\big\|_{\text{op}} > \frac{r \wedge 1}{2} \Big\}.$$

Therefore, with $r = \frac{a}{\mathbb{E}[X_0^2]}$ and using Proposition 3

$$\begin{aligned}
\mathbb{P}(\Gamma_r^c) &\leq \mathbb{P}\left(\big\|\widehat{\Sigma}_m - \Sigma_m\big\|_{\text{op}} > \frac{r \wedge 1}{2 \big\|\Sigma_m^{-1}\big\|_{\text{op}}} \right) \\
&\leq \mathbb{P}\left(\big\|\widehat{\Sigma}_m - \Sigma_m\big\|_{\text{op}} > \frac{a\,(r \wedge 1)}{2} \right) \\
&\leq \mathbb{P}\left(\big\|\widehat{\Sigma}_m^* - \Sigma_m\big\|_{\text{op}} > \frac{a\,(r \wedge 1)}{4} \right) + \mathbb{P}\left(\big\|\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\big\|_{\text{op}} > \frac{a\,(r \wedge 1)}{4} \right) \\
&=: \mathbb{P}_1 + \mathbb{P}_2.
\end{aligned}$$

First using Lemma 3 with $u = \frac{a\,(r \wedge 1)}{4}$ and by virtue of **A5**, it follows

$$\begin{aligned}
\mathbb{P}_1 &\leq 2 \exp\Big(-3 \log n \Big) \\
&\leq \frac{2}{n^3}.
\end{aligned}$$

Now let bound $\mathbb{P}_2$. We know that for a $D_m \times D_m$ matrix A

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

Thus, from Markov's Inequality,

$$\begin{aligned}
\mathbb{P}_2 &\leq \frac{4}{a\,(r \wedge 1)} \mathbb{E}\Big[\big\|\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\big\|_{\text{op}} \Big] \\
&\leq \frac{4}{a\,(r \wedge 1)} \mathbb{E}\Big[\max_{1 \leq i \leq D_m} \sum_{j=1}^{D_m} \big| \big(\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\big)_{i,j} \big| \Big] \\
&\leq \frac{4}{a\,(r \wedge 1)} \sum_{j=1}^{D_m} \mathbb{E}\big[\big| \big(\widehat{\Sigma}_m - \widehat{\Sigma}_m^*\big)_{i_0,j} \big| \big] \\
&\leq \frac{4}{a\,(r \wedge 1)} \sum_{j=1}^{D_m} \mathbb{E}\big[\big| X_{t-i_0} X_{t-j} - X_{t-i_0}^* X_{t-j}^* \big| \big].
\end{aligned}$$

Moreover, $\big|X_{t-i}X_{t-j} - X_{t-i}^* X_{t-j}^*\big| \leq \big|X_{t-i}\big|\big|X_{t-j} - X_{t-j}^*\big| + \big|X_{t-j}^*\big|\big|X_{t-i} - X_{t-i}^*\big|$ so that with Cauchy-Schwartz's Inequality,

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

Hence,

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

where the last inequality follows from Proposition 3 and Proposition 1. As a result, choosing $s_n = O(\frac{\sqrt{n}}{\log n})$ (ensuring **A5**), one can find a constant A such that

$$\Big(\frac{\log q_n}{q_n}\Big)^{\gamma-1} \leq A\,\Big(\frac{1}{\sqrt{n}}\Big)^{\gamma-1}. \tag{5.1}$$

As a result, with $\gamma \geq 8$ and $C = 8\,\big\|X_0\big\|_2/(a\,(r \wedge 1))$

$$\mathbb{P}_2 \;\leq\; A\,C\,C_\tau\,D_m\,\frac{1}{n^{7/2}} \;\leq\; A\,C\,C_\tau\,C_K\,\frac{1}{n^3}$$

by virtue of **A6**. The result is proved with $c_0 = A\,C\,C_\tau\,C_K + 1$. ■

5.5 Proof of Proposition 3

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 to the largest eigenvalue $\lambda_{\max}(\Sigma_m)$. Hence,

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

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

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

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

Given that Σ_m is symmetric, it follows

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

which concludes the proof of (2.10).

Now we end by the proof of (2.11). 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

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

so that (2.11) is established.

■

5.6 Technical Lemmas

Lemma 1. *Assume **A1** holds and (X_t) the mixing stationary solution of (1.1). Then, the process $(\vec{X}_t)$ is mixing and*

$$\tau^{(1)}_{\vec{X},\infty}(r) \leq K_n\, \tau^{(1)}_{X,\infty}(r-1). \tag{5.2}$$

Proof. Let set by $\mathcal{M}^i_{\vec{X}} = \sigma(\vec{X}_t, t \leq i)$ and $\mathcal{M}^i_X = \sigma(X_t, t \leq i)$ for an integer i. One would like to bound $\tau\big(\mathcal{M}^i_{\vec{X}}, \big(\vec{X}_{j_1}, \ldots, \vec{X}_{j_k}\big)\big)$ for $j_k > \ldots > j_1 \geq i + r$.
Let assume that the universe Ω is rich enough so that, one can find $\vec{X}^*_{j_l} = \big(X^*_{j_l-1}, \ldots, X^*_{j_l-K_n}\big)^\top$ with $l = 1, \ldots, k$ verifying

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

As a result,

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

This fact along with the definition of $\tau^{(1)}_{\vec{X},\infty}(r)$ leads to (5.2).

■

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 $z_j, j = 0, 1, \ldots$, we find that $\phi_0 = 1$ and for $i \geq 1$

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

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 $h \geq h_0$, it holds $|\phi_t| \leq C\,(t+1)^{-\gamma}$ for some constant $C > 0$. Thus,

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

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 $u > 0$*

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

Proof. One can write for a matrix A

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

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

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

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 $\mathrm{SG}(D_m\,\sigma_0^2)$. Therefore, Y_t^2 is $\mathrm{SE}(256\,D_m^2\,\sigma_0^4, 16\,D_m\sigma_0^2)$ (where SE stands for Sub-Gaussian and SE for Sub-Exponential).

Moreover, we can write

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

Therefore,

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

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

$\{\mathbf{Y}_{1,k}\}$ and $\{\mathbf{Y}_{2,k}\}$ are independent random vectors by virtue of Proposition 2. 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 $w_i = Y^2_{2kq_n+i} - \mathbb{E}[Y_1^2]$, we have

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

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

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

so that

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

■

Lemma 4. *For every $m \in \mathcal{M}_n$, it holds*

$$\mathbb{E}\Big[\big| < \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) > \big|\mathbb{1}_{\Gamma_m}\Big] \leq 2\,D_m. \tag{5.3}$$

Proof. We have

$$\begin{aligned} < \xi, \mathbf{P}_{\mathbf{M}_m}(\xi) > &= < \xi, \mathbf{M}_m\big(\mathbf{M}_m^\top\mathbf{M}_m\big)^{-1}\mathbf{M}_m^\top\xi > \\ &= < \mathbf{M}_m^\top\xi, \widehat{\Sigma}_m^{-1}\,\mathbf{M}_m^\top\xi > \\ &= < \mathbf{M}_m^\top\xi, (\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1})\mathbf{M}_m^\top\xi > + < \mathbf{M}_m^\top\xi, \Sigma_m^{-1}\,\mathbf{M}_m^\top\xi > . \end{aligned}$$

On one hand,

$$\mathbb{E}\Big[(\mathbf{M}_m^\top\xi)^\top\,\Sigma_m^{-1}\,(\mathbf{M}_m^\top\xi)\Big] = D_m \tag{5.4}$$

Indeed, let set by $\tilde{\xi} = \mathbf{M}_m^\top\xi$ and for all $k = 1, \ldots, D_m$, $\tilde{\xi}_k = \sum_{t=1}^n X_{t-k}\xi_t$. Using conditional expectation, it can be showed that $\tilde{\xi}$ each component of $\tilde{\xi}$ is a sum of martingale difference sequence. Let compute the covariance matrix of $\tilde{\xi}$. The k,l element of this matrix is

$$\begin{aligned} \big(\Sigma_{\tilde{\xi}}\big)_{k,l} &= \mathbb{E}\big[\sum_{i=1}^n X_{i-k}\xi_i \sum_{j=1}^n X_{j-l}\xi_j\big] \\ &= \mathbb{E}\big[\sum_{i=1}^n X_{i-k}X_{i-l}\big] \\ &= \big(\mathbb{E}[\mathbf{M}_m^\top\mathbf{M}_m]\big)_{k,l} = \big(\Sigma_m\big)_{k,l} \end{aligned}$$

Therefore, $\Sigma_{\tilde{\xi}} = \Sigma_m$ and

$$\mathbb{E}\Big[(\mathbf{M}_m^\top\xi)^\top\,\Sigma_m^{-1}\,(\mathbf{M}_m^\top\xi)\Big] = \mathrm{Trace}\big(\Sigma_m^{-1}\,\Sigma_m\big) = D_m.$$

Moreover, it holds on Γ_m

$$\mathbb{E}\big[| < \mathbf{M}_m^\top\xi, (\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1})\mathbf{M}_m^\top\xi > |\big] \leq D_m. \tag{5.5}$$

Indeed,

$$\begin{aligned} \mathbb{E}\big[| < \mathbf{M}_m^\top\xi, (\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1})\mathbf{M}_m^\top\xi > |\big] &\leq \mathbb{E}\Big[\big\|\tilde{\xi}\big\|\,\big\|(\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1})\,\tilde{\xi}\big\|\Big] \\ &\leq \mathbb{E}\Big[\big\|\tilde{\xi}\big\|\,\big\|\widehat{\Sigma}_m^{-1} - \Sigma_m^{-1}\big\|_{\mathrm{op}}\,\big\|\tilde{\xi}\big\|\Big] \\ &\leq r\,\big\|\Sigma_m^{-1}\big\|_{\mathrm{op}}\,\mathbb{E}\Big[\big\|\tilde{\xi}\big\|^2\Big] \\ &\leq r\,D_m\frac{\mathbb{E}[X_0^2]}{a}, \end{aligned}$$

where the last inequality holds since $\mathbb{E}\big[\|\tilde{\xi}\|^2\big] = \mathrm{Trace}(\Sigma_m) = D_m\,\mathbb{E}[X_0^2]$. This implies (5.5) as $r = \frac{a}{\mathbb{E}[X_0^2]}$. (5.5) and (5.4) lead to (5.3). ■

Lemma 5. *For every $m \in \mathcal{M}_n$, it holds*

$$\mathbb{E}\big[\, < \xi, \big(\mathbf{I}_n - \mathbf{P}_{\mathbf{M}_m}\big)\big(F_{\theta^*_m} - F_{\theta^*}\big) > \,\big] = 0 \tag{5.6}$$

Proof. First, we have

$$\mathbb{E}\big[\, < \xi, \mathbf{I}_n\big(F_{\theta^*_m} - F_{\theta^*}\big) > \,\big] = 0. \tag{5.7}$$

Indeed,

$$\begin{aligned}\mathbb{E}\big[\, < \xi, \mathbf{I}_n\big(F_{\theta^*_m} - F_{\theta^*}\big) > \,\big] &= \sum_{t=1}^{n} \mathbb{E}\big[\xi_t \,(f^t_{\theta^*_m} - f^t_{\theta^*})\big] \\ &= \sum_{t=1}^{n} \mathbb{E}\big[(f^t_{\theta^*_m} - f^t_{\theta^*})\,\mathbb{E}[\xi_t|\mathcal{F}_t]\,\big] = 0.\end{aligned}$$

Secondly, $\mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*_m} - F_{\theta^*}\big)$ is an element of S_m. Therefore, there exists $\theta_0 \in \Theta_m$ possibly dependent on $(X_1, X_2, \ldots, X_n)$ such that

$$\mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*_m} - F_{\theta^*}\big) = \mathbf{M}_m\,\theta_0.$$

As a result,

$$\big|\, < \xi, \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*_m} - F_{\theta^*}\big) > \big| = \big|\, < \xi, \mathbf{M}_m\,\theta_0 > \big| \leq \sup_{\theta\in\Theta_m} \big|\, < \xi, \mathbf{M}_m\,\theta > \big|.$$

Since, $\theta \mapsto \big|\, < \xi, \mathbf{M}_m\,\theta > \big|$ is a continuous function and Θ_m compact, one can find $\theta_1 \in \Theta_m$ such that

$$\sup_{\theta\in\Theta_m} \big|\, < \xi, \mathbf{M}_m\,\theta > \big| = \big|\, < \xi, \mathbf{M}_m\,\theta_1 > \big|.$$

But, for any $\theta \in \Theta_m$, $\mathbb{E}\big[|\, < \mathbf{M}_m^\top\,\xi, \theta > |\big] \leq \sum_{k=1}^{D_m} \mathbb{E}\big[|\theta_k\,\widetilde{\xi}_k|\big] = 0$. It then follows

$$\mathbb{E}\big[|\, < \xi, \mathbf{P}_{\mathbf{M}_m}\big(F_{\theta^*_m} - F_{\theta^*}\big) > |\big] \leq \mathbb{E}\big[|\, < \mathbf{M}_m^\top\,\xi, \theta_1 > |\big] = 0,$$

which along with (5.7) implies (5.6). ■

Lemma 6. *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 $\mathbf{M}_m = \big[X_{i-1}, \ldots, X_{i-D_m}\big]_{i=1}^n$. By virtue of **A3**, $\mathbf{M}_m$ is of full rank which implies the a.e. invertibility of $\widehat{\Sigma}_m$.

Moreover, $\Sigma_m = \mathbb{E}\big[\widehat{\Sigma}_m\big] = \mathbb{E}\big[Z_0^m(Z_0^m)^\top\big]$ with $Z_0^m = (X_{-1}, \ldots, X_{-D_m})^\top$. Let $\mathbf{u} \in \mathbb{R}^{D_m}$, it follows $\mathbf{u}^\top\Sigma_m\mathbf{u} = \mathbb{E}\big[((Z_0^m)^\top\mathbf{u})^2\big] \geq 0$. 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. ■

6 Acknowledgements

The author thanks William KENGNE Jean-Marc BARDET for proofreads and helpful discussions.

References

[1] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems*, 22:46–54, 2009.

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

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

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

[5] L. Birgé and P. Massart. A generalized cp criterion for gaussian model selection. technical report, universités de paris 6 et paris 7, 2010. prépublication 647,39 pages. 2001.

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

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

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

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

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

[11] A. Goldenshluger and A. Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *Annals of statistics*, pages 417–444, 2001.

[12] D. Hsu, S. M. Kakade, and T. Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.

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

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

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

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

[17] R. Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, pages 147–164, 1980.

[18] R. Shibata. Consistency of model selection and parameter estimation. *Journal of Applied Probability*, pages 127–141, 1986.

[19] S. A. van de Geer. On hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Springer, 2002.