Consistent model selection criteria and goodness-of-fit test for affine causal processes

Jean-Marc Bardet, Kare Kamila, William Kengne

Electronic Journal of Statistics
ISSN: 1935-7524

Consistent model selection criteria and goodness-of-fit test for affine causal processes

Jean-Marc Bardet and Kare Kamila[*]

S.A.M.M., Université Paris 1, Panthéon-Sorbonne,
90, rue de Tolbiac, 75634, Paris, France
e-mail: Jean-Marc.Bardet@univ-paris1.fr and kamilakare@gmail.com

William Kengne

THEMA, Université de Cergy-Pontoise, FRANCE.
e-mail: william.kengne@gmail.com

Abstract: This paper studies the model selection problem in a large class of causal time series models, which includes both the ARMA or AR(∞) processes, as well as the GARCH or ARCH(∞), APARCH, ARMA-GARCH and many others processes. To tackle this issue, we consider a penalized contrast based on the quasi-likelihood of the model. We provide sufficient conditions for the penalty term to ensure the consistency of the proposed procedure as well as the consistency and the asymptotic normality of the quasi-maximum likelihood estimator of the chosen model. It appears from these conditions that the Bayesian Information Criterion (BIC) does not always guarantee the consistency. We also propose a tool for diagnosing the goodness-of-fit of the chosen model based on the portmanteau Test. Numerical simulations and an illustrative example on the FTSE index are performed to highlight the obtained asymptotic results, including a numerical evidence of the non consistency of the usual BIC penalty for order selection of an AR(p) models with ARCH(∞) errors.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: model selection, affine causal processes, consistency, BIC, Portmanteau Test.

1. Introduction

Model selection is an important tool for statisticians and all those who process data. This issue has received considerable attention in the recent literature. There are several model selection procedures, the main ones are : cross validation and penalized contrast based.

The cross validation ([43], [2]) consists in splitting the data into learning sample, which will be used for computing estimators of the parameters and the test sample which allows to assess these estimators by evaluate their risks.

The procedures using penalized objective function search for a model, minimizing a trade-off between a sum of an empirical risk (for instance least squares, $-2\times$log-likelihood), which indicates how well the model fits the data, and a measure of model's complexity so-called a penalty.
The idea of penalizing dates back to the 1970s with the works of [34] and [1]. By using the ordinary least squares in regression framework, Mallows obtained the C_p criterion. Meanwhile, Akaike derived AIC for density estimation using log-likelihood contrast. A few years later, following Akaike, [38] proposed an alternative approach to density estimation and derived the Bayesian Information Criteria (BIC). The penalty term of these criteria is proportional to the dimension of the model. In the recent decades, different approaches of penalization have emerged such as the $\mathbb{L}^2$ norm for the Ridge penalisation [18], the $\mathbb{L}^1$ norm used by [45] that provides the LASSO procedure and the elastic-net that mixes the $\mathbb{L}^1$ and $\mathbb{L}^2$ norms [50].

Model selection procedures can have two different objectives: *consistency* and *efficiency*. A procedure is said to be consistent if given a family of models, including the "true model", the probability of choosing the correct model approaches one as the sample size tends to infinity. On the other hand, a procedure is efficient when its risk is asymptotically equivalent to the risk of the oracle. In this work,

[*]Developed within the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754362.

[†]Developed within the ANR BREAKRISK : ANR-17-CE26-0001-01.

we are interested to construct a consistent procedure for the general class of times series known as *affine causal processes*.

This class of affine causal time series can be defined as follows. Let $\mathbb{R}^\infty$ be the space of sequences of real numbers with a finite number of non zero, if $M,\ f : \mathbb{R}^\infty \to \mathbb{R}$ are two measurable functions, then an affine causal class is

Class $\mathcal{AC}(M,f)$: A process $X = (X_t)_{t\in\mathbb{Z}}$ belongs to $\mathcal{AC}(M,f)$ if it satisfies:

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

where $(\xi)_{t\in\mathbb{Z}}$ is a sequence of zero-mean independent identically distributed random vectors (i.i.d.r.v) satisfying $\mathbb{E}(|\xi_0|^r)<\infty$ for some $r\geq 2$ and $\mathbb{E}[\xi_0^2]=1$.

For instance,

- if $M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\sigma$ and $f\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\phi_1 X_{t-1}+\cdots+\phi_p X_{t-p}$, then $(X_t)_{t\in\mathbb{Z}}$ is an AR(p) process;
- if $M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\sqrt{a_0+a_1X_{t-1}^2+\cdots+a_pX_{t-p}^2}$ and $f\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=0$, then $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process.

Note that, numerous classical time series models such as ARMA(p,q), GARCH(p,q), ARMA(p,q)-GARCH(p,q) (see [12] and [33]) or APARCH(δ,p,q) processes (see [12]) belongs to $\mathcal{AC}(M,f)$. The existence of stationary and ergodic solutions of this class has been studied in [13] and [7].

We consider a trajectory $(X_1,\ldots,X_n)$ of a stationary affine causal process $\mathcal{AC}(M^*,f^*)$, where M^* and f^* are unknown. We also consider a finite set $\mathcal{M}$ of parametric models m, which are affine causal time series. We assume that the "true" model m^* corresponds to M^* and f^*. The aim is to obtain an estimator $\widehat{m}$ of m^* and testing the goodness-of-fit of the chosen model.

There already exist several important contributions devoted to the model selection for time series ; we refer to the book of [35] and the references therein for an overview on this topic.
As we have pointed above, two properties are often used to evaluate a quality of a model selection procedure : consistency and efficiency. The first measure is often used when the true model is included in the collection of model's candidate ; otherwise, efficiency is the well-defined property. In many research in this framework, the main goal is to develop a procedure that fulfills one of these properties. So, in some classical linear time series models, the consistency of the BIC procedure has been established, see for instance [17] or [46] ; and the asymptotic efficiency of the AIC has been proved, see, among others, [41], [20] for a corrected version of AIC for small samples, [23], [21], [22] for the case of infinite order autoregressive model. [40] propose the (consistent) residual information criteria (RIC) for regression model (including regression models with ARMA errors) selection. In the framework of nonlinear threshold models, [25] proved consistency results of a large class of information criteria, whereas [16] focussed on cross-validation type procedure for model selection in a class of semiparametric time series regression model. Let us recall that, the time series model selection literature is very extensive and still growing ; we refer to the monograph of [36], which provided an excellent summary of existing model selection procedure, including the case of time series models as well as the recent review paper of [11].

The adaptive lasso, introduced by [49] for variable selection in linear regression models has been extended by [37] to vector autoregressive models, [26] carried out this procedure in stationary and nonstationary autoregressive models ; the oracle efficient is established. [28] considers model selection for density estimation under mixing conditions and derived oracle inequalities of the slope heuristic procedure ([9] or [5]) ; whereas [3] develop oracle inequalities for model selection for weakly dependent time series forecasting. Recently, [39] have considered the model selection for ARMA time series with trend, and proved the consistency of BIC for the detrended residual sequence, while [4] developed oracle inequalities of sequential model selection method for nonparametric autoregression. [19] pointed out that most existing model selection procedure cannot simultaneously enjoy consistency and (asymptotic) efficiency. They propose a misspecification-resistant information criterion that can achieve consistency and asymptotic efficiency for prediction using model selection.

In this paper, we focus on the class of models (1.1), and addressed the following questions :

1. What regularity conditions are sufficient to build a consistent model selection procedure? Does the classic criterion such as BIC, still have consistent property for choosing a model among the collections $\mathcal{M}$?
2. How can we test the goodness-of-fit of the chosen model?

These questions have not yet been answered for the class of models and the framework considered here, in particular in case of infinite memory processes. This new contribution provides theoretical and numerical response of these issues.

(i) The estimator $\widehat{m}$ of m^* is chosen by minimizing a penalized criterion $\widehat{C}(m) = -2\widehat{L}_n(m) + |m|\,\kappa_n$, where $\widehat{L}_n(m)$ is a Gaussian quasi-log-likelihood of the model m, $|m|$ is the number of estimated parameters of the model m and κ_n is a non-decreasing sequence of real numbers (see more details in Section 2). Note that, in the cases $\kappa_n = 2$ or $\kappa_n = \log n$ we respectively consider the usual AIC and BIC criteria. We provide sufficient conditions (essentially depending on the decreasing of the Lipschitz coefficients of the functions f and M) for obtaining consistency of the model selection procedure. We also theoretically and numerically exhibit an example of order selection (weak AR(p) processes with ARCH(∞) errors) such that the consistency of the classical BIC penalty is not ensured.

(ii) We provide an asymptotic goodness-of-fit test for the selected model that is very simple to be used (with the usual Chi-square distribution limit), which successively completes the model selection procedure. Numerical applications show the accuracy of this test under the null hypothesis as well as an efficient test power under an alternative hypothesis. Note that, similar test has been proposed by [31] under the Gaussian assumption on the observations, whereas [32] focused for multivariate time series with multivariate ARCH-type errors. Also, [14] proposed a portmanteau test statistic based on generalized inverses and $\{2\}$-inverses for diagnostic checking in the class of model (1.1). Unlike these authors, we apply the test to a model obtained from a model selection procedure.

The paper is organized as follows. Some definitions, notations and assumptions are described in Section 2. The consistency of the criteria and the asymptotic normality of the post-model-selection estimator are studied in Section 3. In Section 4, the examples of $AR(\infty)$, $ARCH(\infty)$, $APARCH(\delta, p, q)$ and ARMA(p, q)-GARCH(p', q') processes are detailed. The goodness-of-fit test is presented in Section 5. Finally, numerical results are presented in Section 6 and Section 7 contains the proofs.

2. Definitions and Assumptions

Let us introduce some definitions and assumptions in order to facilitate the presentation.

2.1. Notation and assumptions

In the sequel, we will consider a subset Θ of $\mathbb{R}^d$ ($d \in \mathbb{N}$). We will use the following norms:

- $\|.\|$ denotes the usual Euclidean norm on $\mathbb{R}^\nu$, with $\nu \geq 1$;
- if X is $\mathbb{R}^\nu$-random variable with $r \geq 1$ order moment, we set $\|X\|_r = \big(\mathbb{E}(\|X\|^r)\big)^{1/r}$;
- for any set $\Theta \subseteq \mathbb{R}^d$ and for any $g : \Theta \to \mathbb{R}^{d'}$, $d' \geq 1$, denote $\|g\|_\Theta = \sup\limits_{\theta \in \Theta}\big\{\|g(\theta)\|\big\}$.

In the introduction, to be more concise, we have presented the problem of time series model selection in a very general form. In reality, we will limit our field of study a little bit by considering a semi-parametric framework. Hence, let $(f_\theta)_{\theta \in \Theta}$ and $(M_\theta)_{\theta \in \Theta}$ be two families of known functions such as for any $\theta \in \Theta$, both f_θ, M_θ with real values defined on $\mathbb{R}^\infty$.

We begin by giving a condition on f_θ and M_θ which ensure the existence of a r-order moment, stationary and ergodic time series belonging to $\mathcal{AC}(M_\theta, f_\theta)$. This condition, initially obtained in [13], is written in terms of Lipschitz coefficients of both these functions. Hence, for $\Psi_\theta = f_\theta$ or M_θ, define:

Assumption A(Ψ_θ,Θ): *Assume that* $\|\Psi_\theta(0)\|_\Theta<\infty$ *and there exists a sequence of non-negative real numbers* $\big(\alpha_k(\Psi_\theta,\Theta)\big)_{k\geq 1}$ *such that* $\sum_{k=1}^\infty \alpha_k(\Psi_\theta,\Theta)<\infty$ *satisfying:*

$$\|\Psi_\theta(x)-\Psi_\theta(y)\|_\Theta \leq \sum_{k=1}^\infty \alpha_k(\Psi_\theta,\Theta)|x_k-y_k| \text{ for all } x,y\in\mathbb{R}^\infty.$$

Now for $r\geq 1$, where $\|\xi_0\|_r<\infty$, define:

$$\Theta(r)=\Big\{\theta\in\mathbb{R}^d,\ A(f_\theta,\{\theta\}) \text{ and } A(M_\theta,\{\theta\}) \text{ hold with} \\ \sum_{k=1}^\infty \alpha_k(f_\theta,\{\theta\})+\|\xi_0\|_r\sum_{k=1}^\infty \alpha_k(M_\theta,\{\theta\})<1\Big\}. \quad (2.1)$$

Then, for any $\theta\in\Theta(r)$, there exists a stationary and ergodic solution with r-order moment belonging to $\mathcal{AC}(M_\theta,f_\theta)$. (see [13] and [7]).

2.2. The framework

Let us start with an example to better understand the framework and the approach of model selection we will follow.

Example: Assume that the observed trajectory $(X_1,\ldots,X_n)$ is generated from an AR(2) process and we would like to identify this family of process and its order. Then, we consider the collection $\mathcal{M}$ of ARMA(p,q) and GARCH(p',q') processes for $0\leq p,q,p',q'\leq 9$ and we would like to chose in this family a "best" model for fitting $(X_1,\ldots,X_n)$. Note that there is 200 possible models and we expect to recognize the AR(2) as the selected model, at least when n is large enough.

We begin with the following property that allow to enlarge the family of models by extending the dimension d of the parameter θ:

Proposition 1. *Let* $d_1,d_2\in\mathbb{N}$, $\Theta_1\subset\mathbb{R}^{d_1}$ *and* $\Theta_2\subset\mathbb{R}^{d_2}$, *and for* $i=1,2$, *define* $f^{(i)}_{\theta_i},M^{(i)}_{\theta_i}:\mathbb{R}^\infty\to\mathbb{R}$ *and for* $\theta_i\in\Theta_i$. *Then there exist* $\max(d_1,d_2)\leq d\leq d_1+d_2$, $\Theta\subset\mathbb{R}^d$, *and a family of functions* $f_\theta:\mathbb{R}^\infty\to\mathbb{R}$ *and* $M_\theta:\mathbb{R}^\infty\to[0,\infty)$ *with* $\theta\in\Theta$, *such that for any* $\theta_1\in\Theta_1$ *and* $\theta_2\in\Theta_2$, *there exists* $\theta\in\Theta$ *satisfying*

$$\mathcal{AC}\big(M^{(1)}_{\theta_1},f^{(1)}_{\theta_1}\big)\bigcup\mathcal{AC}\big(M^{(2)}_{\theta_2},f^{(2)}_{\theta_2}\big)\subset\mathcal{AC}\big(M_\theta,f_\theta\big).$$

The proof of this proposition, as well as the other proofs, can be found in Section 7. This proposition says that it is always possible to embed two parametric causal affine models in a larger one. Hence, for instance, we can consider as well AR processes and ARCH processes in a unique representation, *i.e.*

$$\begin{cases} AR & \begin{cases} M^{(1)}_{\theta_1}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\sigma \\ f^{(1)}_{\theta_1}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\phi_1X_{t-1}+\cdots+\phi_pX_{t-p}\end{cases} \\ ARCH & \begin{cases} M^{(2)}_{\theta_2}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\sqrt{a_0+a_1X^2_{t-1}+\cdots+a_qX^2_{t-q}} \\ f^{(2)}_{\theta_2}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=0\end{cases}\end{cases}$$

$$\Longrightarrow \begin{cases} M_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\sqrt{\theta_0+\theta_1X^2_{t-1}+\cdots+\theta_qX^2_{t-q}} \\ f_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big)=\theta_{q+1}X_{t-1}+\cdots+\theta_{q+p}X_{t-p}\end{cases}.$$

From now and in all the sequel, we fix $d\in\mathbb{N}^*$, and the family of functions $f_\theta,M_\theta:\mathbb{R}^\infty\to\mathbb{R}$ for $\theta\in\Theta\subset\Theta(r)\subset\mathbb{R}^d$.

Let $(X_1,\ldots,X_n)$ be an observed trajectory of an affine causal process X belonging to $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$, where θ^* is an unknown vector of Θ, and therefore:

$$X_t=M_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t+f_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \text{ for any } t\in\mathbb{Z}. \quad (2.2)$$

In the sequel, we will consider several models, which all are particular cases of $\mathcal{AC}(M_\theta, f_\theta)$ with $\theta \in \Theta \subset \mathbb{R}^d$. More precisely define:

- a model m as a subset of $\{1, \ldots, d\}$ and denote $|m| = \#(m)$;
- $\Theta(m) = \big\{(\theta_i)_{1 \leq i \leq d} \in \mathbb{R}^d,\ \theta_i = 0 \text{ if } i \notin m\big\} \cap \Theta$;
- $\mathcal{M}$ as a family of models, *i.e.* $\mathcal{M} \subset \mathcal{P}\big(\{1, \ldots, d\}\big)$.

Finally, for all $m \in \mathcal{M}$, $m \in \mathcal{AC}(M_\theta, f_\theta)$ when $\theta \in \Theta(m)$ and denote m^* the "true" model. We could as well consider hierarchical or exhaustive families of models.

Example: From the previous example, we can consider:
• a family $\mathcal{M}_1$ of models m_1 such as $\mathcal{M}_1 = \big\{\{1\}, \{1,2\}, \ldots, \{1, \ldots, q+1\}\big\}$: this family is the hierarchical one of ARCH processes with orders varying from 0 to q.
• a family $\mathcal{M}_2$ of models m_2 such as $\mathcal{M}_2 = \mathcal{P}\big(\{1, \ldots, p+q+1\}\big)$: this family is the exhaustive one and contains as well the AR(2) process $X_t = \phi_2 X_{t-2} + \theta_0\, \xi_t$ as the process $X_t = \phi_1 X_{t-1} + \phi_3 X_{t-3} + \xi_t \sqrt{\theta_0 + a_2 X_{t-2}^2}$.

To establish the consistency of the selected model, we will need to assume that the "true" model m^* with the parameter θ^*, is included in the model family $\mathcal{M}$.

2.3. The special case of NLARCH(∞) processes

As in [7], in the special case of NLARCH(∞) processes, including for instance GARCH(p, q) or ARCH(∞) processes, a particular treatment can be realized for obtaining sharper results than using the previous framework. In such case, define the class:

Class $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$: A process $X = (X_t)_{t \in \mathbb{Z}}$ belongs to $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$ if it satisfies:

$$X_t = \xi_t \sqrt{\widetilde{H}_\theta\big((X_{t-i}^2)_{i \in \mathbb{N}^*}\big)} \quad \text{for any } t \in \mathbb{Z}. \tag{2.3}$$

Therefore, if $M_\theta^2\big((X_{t-i})_{i \in \mathbb{N}^*}\big) = H_\theta\big((X_{t-i})_{i \in \mathbb{N}^*}\big) = \widetilde{H}_\theta\big((X_{t-i}^2)_{i \in \mathbb{N}^*}\big)$ then, $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta) = \mathcal{AC}(M_\theta, 0)$. In case of the class $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$, we will use the assumption $A(\widetilde{H}_\theta, \Theta)$. By this way, we will obtain a new set of stationary solutions. For $r \geq 2$ define:

$$\widetilde{\Theta}(r) = \Big\{\theta \in \mathbb{R}^d,\ A(\widetilde{H}_\theta, \{\theta\}) \text{ holds with } \big(\|\xi_0\|_r\big)^2 \sum_{k=1}^{\infty} \alpha_k(\widetilde{H}_\theta, \{\theta\}) < 1\Big\}. \tag{2.4}$$

Then, for $\theta \in \Theta(r)$, a process $(X_t)_{t \in \mathbb{Z}}$ belonging to the class $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$ is stationary ergodic and satisfies $\|X_0\|_r < \infty$.

2.4. The Gaussian quasi-maximum likelihood estimation and the model selection criterion

In the sequel, for a model $m \in \mathcal{M}$, a family of models of $\mathcal{AC}(M_\theta, f_\theta)$ with $\theta \in \Theta \subset \mathbb{R}^d$, where $\theta \to M_\theta$ and $\theta \to f_\theta$ are two fixed functions, we are going to consider Gaussian quasi-maximum likelihood estimators (QMLE) of θ for each specific model m.

This approach as semi-parametric estimation has been successively introduced for GARCH(p, q) processes in [24] where its consistency is also proved, and the asymptotic normality of this estimator has been established in [8] and [15]. In [7], those results have been extended to affine causal processes, and an extension to Laplacian QMLE has been also proposed in [6].
The Gaussian QMLE is derived from the conditional (with respect to the filtration $\sigma\big\{(X_t)_{t \leq 0}\big\}$) log-likelihood of $(X_1, \ldots, X_n)$ when (ξ_t) is supposed to be a Gaussian standard white noise. Due to the linearity of a causal affine process, we deduce that this conditional log-likelihood (up to an additional constant) L_n is defined for all $\theta \in \Theta$ by:

$$L_n(\theta) := -\frac{1}{2} \sum_{t=1}^{n} q_t(\theta)\ , \ \text{with } q_t(\theta) := \frac{(X_t - f_\theta^t)^2}{H_\theta^t} + \log(H_\theta^t) \tag{2.5}$$

where $f_\theta^t := f_\theta(X_{t-1}, X_{t-2}, \cdots)$, $M_\theta^t := M_\theta(X_{t-1}, X_{t-2}, \cdots)$ and $H_\theta^t = \big(M_\theta^t\big)^2$. Since $L_n(\theta)$ depends on $(X_t)_{t\leq 0}$ that are unknown, the idea of the quasi log-likelihood is to replace $q_t(\theta)$ by an approximation $\widehat{q}_t(\theta)$ and to compute $\widehat{\theta}$ as in equation (2.7) even if the white noise is not Gaussian. Hence, the conditional quasi log-likelihood (up to an additional constant) is given for all $\theta \in \Theta$ by

$$\widehat{L}_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} \widehat{q}_t(\theta)\ , \ \text{with } \widehat{q}_t(\theta) := \frac{(X_t - \widehat{f}_\theta^t)^2}{\widehat{H}_\theta^t} + \log(\widehat{H}_\theta^t)$$

$$\text{where} \left\{ \begin{array}{lcl} \widehat{f}_\theta^t & := & f_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{M}_\theta^t & := & M_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{H}_\theta^t & := & (\widehat{M}_\theta^t)^2 \end{array} \right. \tag{2.6}$$

for any deterministic sequence $u = (u_n)$ with finitely many non-zero values ($u = 0$ is very often chosen without loss of generality).

However, the definitions of the conditional log-likelihood and quasi log-likelihood require that their denominators do not vanish. Hence, we will suppose in the sequel that the lower bound of $H_\theta(\cdot) = \big(M_\theta(\cdot)\big)^2$ (which is reached since Θ is compact) is strictly positive:

Assumption D(Θ): *$\exists \underline{h} > 0$ such that $\inf_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ for all $x \in \mathbb{R}^\infty$.*

Finally, under this assumption, for each specific model $m \in \mathcal{M}$, we define the Gaussian QMLE $\widehat{\theta}(m)$ as

$$\widehat{\theta}(m) = \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\ \widehat{L}_n(\theta). \tag{2.7}$$

To select the "best" model $m \in \mathcal{M}$, we chose a penalized contrast $\widehat{C}(m)$ ensuring a trade-off between -2 times the maximized quasi log-likelihood, which decreases with the size of the model, and a penalty increasing with the size of the model. Therefore, the choice of the "best" model $\widehat{m}$ among the estimated can be performed by minimizing the following criteria

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \widehat{C}(m) \quad \text{with} \quad \widehat{C}(m) = -2\widehat{L}_n\big(\widehat{\theta}(m)\big) + |m|\,\kappa_n, \tag{2.8}$$

where

- $(\kappa_n)_n$ an increasing sequence depending on the number of observations n.
- $|m|$ denotes the dimension of the model m, *i.e.* the cardinal of m, subset of $\{1, \ldots, d\}$, which is also the number of estimated components of θ (the others are fixed to zero).

The consistency of the criterion $\widehat{C}$, *i.e.*

$$\mathbb{P}(\widehat{m} = m^*) \underset{n\to\infty}{\longrightarrow} 1; \tag{2.9}$$

will be established after showing that both of following probabilities are zero:

- the asymptotic probability of selecting a larger model containing the true model (overfitting case);
- the asymptotic probability of selecting a false model that is a model not containing m^*.

3. Asymptotic results

3.1. Assumptions required for the asymptotic study

The following classical assumption ensures the identifiability of the model considered.

Assumption Id(Θ): *For all θ, $\theta' \in \Theta$, $(f_\theta^0 = f_{\theta'}^0$ and $M_\theta^0 = M_{\theta'}^0)$ a.s. $\Longrightarrow \theta = \theta'$.*

Another required assumption concerns the differentiability of $\Psi_\theta = f_\theta$ or M_θ on Θ. This type of assumption has already been considered in order to apply the QMLE procedure (see [7], [44], [48]). First, the following Assumption Var(Θ) provides the invertibility of the "Fisher's information matrix" of X and is important to prove the asymptotic normality of the QMLE.

Assumption Var: $\big(\sum_{i=1}^d \alpha_i \frac{\partial f_\theta^0}{\partial \theta^{(i)}} = 0 \implies \forall i = 1, \dots, d,\ \alpha_i = 0\ a.s\big)$ *or* $\big(\sum_{i=1}^d \alpha_i \frac{\partial H_\theta^0}{\partial \theta^{(i)}} = 0 \implies \forall i = 1, \dots, d,\ \alpha_i = 0\ a.s\big)$.

Moreover, one of the following technical assumption is required to establish the consistency of the model selection procedure.

Assumption $\boldsymbol{K}(\Theta)$: *Assumptions $A(f_\theta, \Theta), A(M_\theta, \Theta)$, $A(\partial_\theta f_\theta, \Theta)$, $A(\partial_\theta M_\theta, \Theta)$ and $B(\Theta)$ hold and there exists $r \geq 2$ such that $\theta^* \in \Theta(r)$. Moreover, with $s = \min(1, r/3)$, assume that the sequence $(\kappa_n)_{n\in\mathbb{N}}$ satisfies*

$$\sum_{k\geq 1} \big(\frac{1}{\kappa_k}\big)^s \Big(\sum_{j\geq k} \alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta)\Big)^s < \infty.$$

Assumption $\widetilde{\boldsymbol{K}}(\Theta)$: *Assumptions $A(\widetilde{H}_\theta, \Theta)$, $A(\partial_\theta \widetilde{H}_\theta, \Theta)$ and $B(\Theta)$ hold and there exists $r \geq 2$ such that $\theta^* \in \Theta(r)$. Moreover, with $s = \min(1, r/4)$, assume that the sequence $(\kappa_n)_{n\in\mathbb{N}}$ satisfies*

$$\sum_{k\geq 1} \big(\frac{1}{\kappa_k}\big)^s \Big(\sum_{j\geq k} \alpha_j(\widetilde{H}_\theta, \Theta) + \alpha_j(\partial_\theta \widetilde{H}_\theta, \Theta)\Big)^s < \infty.$$

Remark 1. These conditions on $(\kappa_n)_{n\in\mathbb{N}}$ have been deduced from conditions for strong law of large numbers obtained in [27] and are not too restrictive: for instance, if the Lipschitzian coefficients of f_θ, M_θ (the case using $\widetilde{H}_\theta$ can be treated similarly) and their derivatives are bounded by a geometric or Riemanian decrease:

1. the geometric case: $\alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta) = O(a^j)$ with $0 \leq a < 1$, then any (κ_n) such as $1/\kappa_n = o(1)$ can be chosen; for instance $\kappa_n = \log n$ or $\log(\log n)$; this is the case for instance of ARMA, GARCH, APARCH or ARMA-GARCH processes.
2. the Riemanian case: $\alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta) = O(j^{-\gamma})$ with $\gamma > 1$:
   - if $r \geq 3$ then
     - if $\gamma > 2$ then any sequence such as $1/\kappa_n = o(1)$ can be chosen;
     - if $1 < \gamma < 2$, any (κ_n) such as $\kappa_n = O(n^\delta)$ with $\delta > 2 - \gamma$ can be chosen.
   - if $1 \leq r < 3$
     - if $\gamma > (r+3)/r$ then any sequence such as $1/\kappa_n = o(1)$ can be chosen;
     - if $1 < \gamma < (r+3)/r$ then any (κ_n) such as $\kappa_n = n^\delta$ with $\delta > (r+3)/r - \gamma$ can be chosen.

   In the last case of these two conditions on r, we can see the usual BIC choice, $\kappa_n = \log n$ does not fulfill the assumption in general.

3.2. New versions of limit theorems in [7]

These assumptions K(Θ) and $\widetilde{K}(\Theta)$ used in Lemmas 1 and 2 (see Section 7) and the detailed Riemanian convergence rates of the previous remark, provide an improvement of the two main limit theorems established in [7]. More precisely, we obtain:

New version of Theorem 1 in [7]
Let $(X_1, \dots, X_n)$ be an observed trajectory of an affine causal process X belonging to $\mathcal{AC}(M_{\theta^*}, f_{\theta^*})$ (or

$\widetilde{\mathcal{AC}}(\widetilde H_\theta)$) where θ^* is an unknown vector of Θ, a compact set included in $\Theta(r) \subset \mathbb{R}^d$ (or $\widetilde\Theta(r) \subset \mathbb{R}^d$) with $r \geq 2$. Then, if assumptions $A(f_\theta,\Theta)$, $A(M_\theta,\Theta)$ (or $A(\widetilde H_\theta,\Theta)$), $D(\Theta)$, $Id(\Theta)$ hold with

$$\left\{\begin{array}{lll} \alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)=O\big(j^{-\ell}\big) & \text{for some} & \ell>\max(1\,,3/r) \\ \text{or} \quad \alpha_j(\widetilde H_\theta,\Theta)=O\big(j^{-\widetilde\ell}\big) & \text{for some} & \widetilde\ell>\max(1\,,4/r) \end{array}\right. , \tag{3.1}$$

then the QMLE $\widehat\theta(m^*)$ satisfies $\widehat\theta(m^*) \xrightarrow[n\to+\infty]{a.s.} \theta^*$.

Proof. We use the same proof as in [7] except for establishing $\frac{1}{n}\big\|\widehat L_n(\theta)-L_n(\theta)\big\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$. Indeed, we can apply Lemma 1 with $\kappa_n=n$. Hence, this is checked under assumption $\boldsymbol{K}(\Theta)$ under Riemanian condition of Remark 1 if $r\geq 3$ when $\gamma=\ell>1$ and if $2\leq r\leq 3$, when $\gamma=\ell>3/r$, implying the first new conditions of the Theorem.
Under assumption $\widetilde{\boldsymbol{K}}(\Theta)$, an adaptation of Remark 1 implies that for $r\geq 4$ we should have $\gamma=\widetilde\ell>1$ and if $2\leq r\leq 4$, when $\gamma=\widetilde\ell>4/r$. ■

Therefore, in all the previous cases and when $r=4$, we obtain a limiting decrease rate $O\big(j^{-\gamma}\big)$ with $\gamma>1$ instead of $\gamma>3/2$ in [7]. This can also be used to improve Theorem 2 in [7]:

New version of Theorem 2 in [7]
If $r\geq 4$ and under the assumptions of the previous new version of Theorem 1 in [7], and $\mathbf{Var}(\Theta)$, *and if assumptions $A(\partial_\theta f_\theta,\Theta)$, $A(\partial_\theta M_\theta,\Theta)$, $A(\partial^2_{\theta^2} f_\theta,\Theta)$ and $A(\partial^2_{\theta^2} M_\theta,\Theta)$ (or $A(\partial_\theta \widetilde H_\theta,\Theta)$ and $A(\partial^2_{\theta^2}\widetilde H_\theta,\Theta)$) hold with*

$$\left\{\begin{array}{c} \alpha_j(\partial_\theta f_\theta,\Theta)+\alpha_j(\partial_\theta M_\theta,\Theta)=O\big(j^{-\ell'}\big) \\ \text{or} \quad \alpha_j(\partial_\theta \widetilde H_\theta,\Theta)=O\big(j^{-\ell'}\big) \end{array}\right. \quad \text{for some} \;\; \ell'>1, \tag{3.2}$$

then the QMLE $\widehat\theta_n(m^*)$ satisfies

$$\sqrt{n}\,\Big(\big(\widehat\theta(m^*)\big)_i-(\theta^*)_i\Big)_{i\in m^*} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_{|m^*|}\big(0\;,\;F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}\big), \tag{3.3}$$

with $\big(F(\theta^*,m^*)\big)_{i,j}=\mathbb{E}\Big[\dfrac{\partial^2 q_0(\theta^*)}{\partial\theta_i\partial\theta_j}\Big]$ and $(G(\theta^*,m^*))_{i,j}=\mathbb{E}\Big[\dfrac{\partial q_0(\theta^*)}{\partial\theta_i}\dfrac{\partial q_0(\theta^*)}{\partial\theta_j}\Big]$ for $i,\,j\in m^*$.

3.3. Asymptotic model selection

Using the above assumptions, we can establish the limit theorem below, which provides sufficient conditions for the consistency of the model selection procedure.

Theorem 3.1. *Let $(X_1,\ldots,X_n)$ be an observed trajectory of an affine causal process X belonging to $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$ (or $\widetilde{\mathcal{AC}}(\widetilde H_\theta)$) where θ^* is an unknown vector of Θ a compact set included in $\Theta(r)\subset\mathbb{R}^d$ (or $\widetilde\Theta(r)\subset\mathbb{R}^d$) with $r\geq 4$. If assumptions $D(\Theta)$, $Id(\Theta)$, $K(\Theta)$ (or $\widetilde K(\Theta)$), $A(\partial^2_{\theta^2} f_\theta,\Theta)$ and $A(\partial^2_{\theta^2} M_\theta,\Theta)$ (or $A(\partial^2_{\theta^2}\widetilde H_\theta,\Theta)$) also hold, then*

$$\mathbb{P}(\widehat m=m^*) \xrightarrow[n\to\infty]{} 1 \quad \textit{and} \quad \widehat\theta(\widehat m) \xrightarrow[n\to\infty]{\mathcal{P}} \theta^*. \tag{3.4}$$

The following theorem shows the asymptotic normality of the QMLE of the chosen model.

Theorem 3.2. *Under the assumptions of Theorem 3.1 and if $\theta^*\in\overset{\circ}{\Theta}$ and* $\mathbf{Var}(\Theta)$ *hold, then*

$$\sqrt{n}\,\Big(\big(\widehat\theta(\widehat m)\big)_i-(\theta^*)_i\Big)_{i\in m^*} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_{|m^*|}\big(0\;,\;F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}\big), \tag{3.5}$$

where F and G are defined in (3.3).

Remark 2. In Remark 1, we detailed some situations where the assumption $K(\Theta)$ (or $\widetilde K(\Theta)$) holds, which leads to the results of Theorem 3.1 and 3.2. In particular, the $\log n$ penalty usually linked to BIC is consistent in the case of a geometric decrease of the Lipschitz coefficients of the functions f_θ and M_θ (and their first order derivative). In the case of a Riemanian rate, the consistency of BIC is not ensured; see also the next section.

4. Examples

In this section, some examples of time series satisfying the conditions of previous results are considered. These examples include $AR(\infty)$, $ARCH(\infty)$, $APARCH(\delta,p,q)$ and ARMA(p,q)-GARCH(p',q').

4.1. $AR(\infty)$ models

For $(\psi_k(\theta))_{k\in\mathbb{N}}$ a sequence of real numbers depending on $\theta\in\mathbb{R}^d$, let us consider an $AR(\infty)$ process defined by:

$$X_t=\sum_{k\geq 1}\psi_k(\theta^*)X_{t-k}+\sigma\,\xi_t \quad \text{for any } t\in\mathbb{Z}, \tag{4.1}$$

where $(\xi_t)_t$ admits 4-order moments, and $\theta^*\in\Theta\subset\Theta(4)$, the set of $\theta\in\mathbb{R}^d$ such that $\sum_{k\geq 1}\|\psi_k(\theta)\|_\Theta<1$ and $\sigma>0$. This process corresponds to (2.2) with $f_\theta\big((x_i)_{i\geq 1}\big)=\sum_{k\geq 1}\psi_k(\theta)x_k$ and $M_\theta\equiv\sigma$ for any $\theta\in\Theta$. The Lipschitz coefficients of f_θ are $\alpha_k(f_\theta)=\|\psi_k(\theta)\|_\Theta$. Moreover, Assumption $D(\Theta)$ holds with $\underline{h}=\sigma^2>0$.

Let us consider $\mathcal{M}$ a finite family of models. Of course, the main example of such family of models is given by the one of ARMA(p,q) processes with $0\leq p\leq p_{\max}$ and $0\leq q\leq q_{\max}$, providing $(p_{\max}+1)(q_{\max}+1)$ models and $\theta\in\mathbb{R}^{p_{\max}+q_{\max}+1}$.

Besides, assume that $Id(\Theta)$, Var(Θ) hold and that the sequence (ψ_k) is twice differentiable (with respect to θ) on Θ, with $\sum_k\|\partial^2_\theta\psi_k(\theta)\|_\Theta<\infty$ and $\|\psi_k(\theta)\|_\Theta+\|\partial_\theta\psi_k(\theta)\|_\Theta=O(k^{-\gamma})$ with $\gamma>1$. From Remark 1,

- if $\gamma>2$, the condition $\kappa_n\underset{n\to\infty}{\longrightarrow}\infty$ (for instance, the BIC penalization with $\kappa_n=\log(n)$, or $\kappa_n=\sqrt{n}$) ensures the consistency of $\widehat{m}$ and the Theorem (3.2) holds if in addition $\theta^*\in\overset{\circ}{\Theta}$;
- if $1<\gamma<2$, $\kappa_n=O(n^\delta)$ with $\delta>2-\gamma$ has to be chosen (and we cannot insure the consistency of $\widehat{m}$ in case of classical BIC penalization).

Finally, in the particular case of the family of ARMA processes, the stationarity condition implies that any $\kappa_n\underset{n\to\infty}{\longrightarrow}\infty$ can be chosen (BIC penalization with $\kappa_n=\log(n)$, or $\kappa_n=\sqrt{n}$), since the decreases of ψ_k and its derivative are exponential.

4.2. $ARCH(\infty)$ models

For $(\psi_k(\theta))_{k\in\mathbb{N}}$ a sequence of nonnegative real numbers depending on $\theta\in\mathbb{R}^d$, with $\psi_0>0$, let us consider an ARCH(∞) process defined by :

$$X_t=\Big(\psi_0(\theta^*)+\sum_{k=1}^{\infty}\psi_k(\theta^*)X^2_{t-k}\Big)^{1/2}\xi_t \quad \text{for any } t\in\mathbb{Z}, \tag{4.2}$$

where $\mathbb{E}\big[\xi_0^4\big]<\infty$, and $\theta^*\in\Theta\subset\widetilde{\Theta}(4)$, the set of $\theta\in\mathbb{R}^d$ such that $\sum_{k\geq 1}\|\psi_k(\theta)\|_\Theta<1$. This process corresponds to (2.2) with $f_\theta\big((x_i)_{i\geq 1}\big)\equiv 0$ and $H_\theta\big((x_i)_{i\geq 1}\big)=\psi_0(\theta)+\sum_{k=1}^\infty\psi_k(\theta)x_k^2$, *i.e.* $\widetilde{H}_\theta\big((y_i)_{i\geq 1}\big)=\psi_0(\theta)+\sum_{k=1}^\infty\psi_k(\theta)y_k$, for any $\theta\in\Theta$. The Lipschitz coefficients of $\widetilde{H}_\theta$ are $\alpha_k(\widetilde{H}_\theta)=\|\psi_k(\theta)\|_\Theta$. Moreover, Assumption $D(\Theta)$ holds if $\underline{h}=\inf_{\theta\in\Theta}\psi_0(\theta)>0$.

Let us consider $\mathcal{M}$ a finite family of models. The main example of such family of models is given by the GARCH(p,q) processes with $0\leq p\leq p_{\max}$ and $0\leq q\leq q_{\max}$, providing $(p_{\max}+1)(q_{\max}+1)$ models and $\theta\in\mathbb{R}^{p_{\max}+q_{\max}+1}$.

Moreover, assume that $Id(\Theta)$, Var(Θ) hold and that the sequence (ψ_k) is twice differentiable (with respect to θ) on Θ, with $\sum_k\|\partial^2_\theta\psi_k(\theta)\|_\Theta<\infty$ and $\|\psi_k(\theta)\|_\Theta+\|\partial_\theta\psi_k(\theta)\|_\Theta=O(k^{-\gamma})$ with $\gamma>1$. From Remark 1,

- if $\gamma > 2$, the condition $\kappa_n \underset{n\to\infty}{\longrightarrow} \infty$ (for instance, the BIC penalization with $\kappa_n = \log(n)$, or $\kappa_n = \sqrt{n}$) ensures the consistency of $\widehat{m}$ and the Theorem (3.2) holds if in addition, $\theta^* \in \overset{\circ}{\Theta}$;
- if $1 < \gamma < 2$, $\kappa_n = O(n^\delta)$ with $\delta > 2 - \gamma$ has to be chosen (and we cannot insure the consistency of $\widehat{m}$ in the case of the classical BIC penalization).

Finally, in the particular case of the family of GARCH processes, the stationarity condition implies that any $\kappa_n \underset{n\to\infty}{\longrightarrow} \infty$ can be chosen (BIC penalization with $\kappa_n = \log(n)$, or $\kappa_n = \sqrt{n}$), since the decreases of ψ_k and its derivative are exponential.

4.3. APARCH(δ, p, q) models

For $\delta \geq 1$ and from [12], $(X_t)_{t\in\mathbb{Z}}$ is an APARCH(δ, p, q) process with $p, q \geq 0$ if:

$$\begin{cases} X_t = \sigma_t\, \xi_t \\ (\sigma_t)^\delta = \omega + \sum_{i=1}^p \alpha_i(|X_{t-i}| - \gamma_i X_{t-i})^\delta + \sum_{j=1}^q \beta_j(\sigma_{t-j})^\delta \quad \text{for any } t \in \mathbb{Z}, \end{cases} \tag{4.3}$$

where $\omega > 0$, $-1 < \gamma_i < 1$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for $1 \leq i \leq p$ and $1 \leq j \leq q$, $\alpha_p > 0$, $\beta_q > 0$ and $\sum_{j=1}^q \beta_j < 1$. From [6], with $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \gamma_1, \ldots, \gamma_p, \beta_1, \ldots, \beta_p)'$, the conditional variance σ_t can be rewritten as follows

$$\sigma_t^\delta = b_0(\theta) + \sum_{k\geq 1} \Big(b_k^+(\theta)(\max(X_{t-k}, 0))^\delta - b_k^-(\theta)(\min(X_{t-k}, 0))^\delta \Big);$$

with $f_\theta \equiv 0$ and $M_\theta^t = \sigma_t$, we deduce that $\alpha_k(M_\theta, \Theta) = \max(\|b_k^+(\theta)\|_\Theta^{1/\delta}, \|b_k^-(\theta)\|_\Theta^{1/\delta})$, and from the assumption $\sum_{j=1}^q \beta_j < 1$, the Lipschitz coefficients $\alpha_k(M_\theta, \Theta)$ decrease exponentially fast. Then, the stationarity set for $r \geq 1$ is

$$\Theta(r) = \Big\{ \theta \in \mathbb{R}^{2p+q+1} \Big/ \|\xi_0\|_r \sum_{j=1}^\infty \max\big(|b_j^+(\theta)|^{1/\delta}, |b_j^-(\theta)|^{1/\delta}\big) < 1 \Big\}.$$

Now, assume that $(X_t)_{t\in\mathbb{Z}}$ is an APARCH(δ, p^*, q^*) where $0 \leq p^* \leq p_{\max}$ and $0 \leq q^* \leq q_{\max}$ are unknown orders as well as the other parameters: $\omega^* > 0$, $-1 < \gamma_i^* < 1$, $\alpha_i^* \geq 0$, $\beta_j^* \geq 0$ for $1 \leq i \leq p_{\max}$ and $1 \leq j \leq q_{\max}$, $\alpha_{p^*} > 0$, $\beta_{q^*} > 0$.

Let $\mathcal{M}$ be the family of APARCH(δ, p, q) processes, with $0 \leq p \leq p_{\max}$ and $0 \leq q \leq q_{\max}$. As a consequence, we consider here $d = 2p_{\max} + q_{\max} + 1$, and

$$\theta^* = {}^t\big(\omega^*, \alpha_1^*, \ldots, \alpha_{p^*}^*, 0, \ldots, 0, \gamma_1^*, \ldots, \gamma_{p^*}^*, 0, \ldots, 0, \beta_1^*, \ldots, \beta_{q^*}^*, 0, \ldots, 0\big) \in \mathbb{R}^d.$$

With all the previous conditions, assumptions D(Θ), Id(Θ), Var(Θ) are satisfied. Moreover, since the Lipschitz coefficients decrease exponentially fast, K(Θ) is satisfied when $\kappa_n \to \infty$. Therefore, the consistency Theorem (3.1) and the Theorem (3.2) of the estimator of the chosen model are satisfied when $r = 4$ and $\kappa_n \to \infty$ (for instance with the typical BIC penalty $\kappa_n = \log n$).

4.4. ARMA(p, q)-GARCH(p', q') models

From [12] and [33], we define $(X_t)_{t\in\mathbb{Z}}$ as an (invertible) ARMA(p, q)-GARCH(p', q') process with $p, q, p', q' \geq 0$ if:

$$\begin{cases} X_t = \sum_{i=1}^p a_i\, X_{t-i} + \varepsilon_t - \sum_{i=1}^q b_i\, \varepsilon_{t-i} \\ \varepsilon_t = \sigma_t\, \xi_t, \ \text{with } \sigma_t^2 = c_0 + \sum_{i=1}^{p'} c_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q'} d_i\, \sigma_{t-i}^2 \end{cases} \quad \text{for all } t \in \mathbb{Z},$$

where

- $c_0 > 0$, $c_{p'} > 0$, $c_i \geq 0$ for $i = 1, \cdots, p' - 1$ and $d_{q'} > 0$, $d_i \geq 0$ for $i = 1, \cdots, q' - 1$;
- $P(x) = 1 - \sum_{i=1}^p a_i x^i$ and $Q(x) = 1 - \sum_{i=1}^q b_i x^i$ are coprime polynomials.

Here we will consider the case of a stationary invertible ARMA(p,q)-GARCH(p',q') process such as $\|X_0\|_4<\infty$ and therefore we will consider:

$$\Theta_{p,q,p',q'}=\Big\{(a_1,\ldots,d_{q'})\in\mathbb{R}^{p+q+p'+1+q'},\ \sum_{j=1}^{q'}d_j+\|\xi_0\|_4\sum_{j=1}^{p'}c_j<1 \\ \text{and } \big(1-\sum_{j=1}^{p}a_jz^j\big)\big(1-\sum_{j=1}^{q}b_jz^j\big)\neq 0 \text{ for all } |z|\leq 1\Big\}.$$

Therefore, if $(a_1,\ldots,d_{q'})\in\Theta_{p,q,p',q'}$, $(\varepsilon_t)_t$ is a stationary GARCH(p',q') process and $(X_t)_t$ is a stationary weak invertible ARMA(p,q) process.

Moreover, following Lemma 2.1. of [6], we know that a stationary ARMA(p,q)-GARCH(p',q') process is a stationary affine causal process with functions f_θ and M_θ satisfying the Assumption A(f_θ,Θ) and A(M_θ,Θ) with Lipschitzian coefficients decreasing exponentially fast, as well as their derivatives. Finally, if Θ is a bounded subset of $\Theta_{p,q,p',q'}$, then assumptions D(Θ), Id(Θ) and Var(Θ) are automatically satisfied.

Assume now that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA(p^*,q^*)-GARCH$(p^{'*},q^{'*})$ process where $0\leq p^*\leq p_{\max}$, $0\leq q^*\leq q_{\max}$, $0\leq p^{'*}\leq p'_{\max}$ and $0\leq q^{'*}\leq q'_{\max}$ are unknown orders with also unknown parameters: $c_0^*,\ldots,c^*_{p'^*},d_1^*,\ldots,d^*_{q'^*},a_1^*,\ldots,a^*_{p^*},b_1^*,\ldots,b_{q^*}$.

Let $\mathcal{M}$ be the family of ARMA(p,q)-GARCH$(p^{'},q^{'})$ processes, with $0\leq p\leq p_{\max}$, $0\leq q\leq q_{\max}$, $0\leq p'\leq p'_{\max}$ and $0\leq q'\leq q'_{\max}$. Hence, we consider here $d=p_{\max}+q_{\max}+p'_{\max}+q'_{\max}+1$, and

$$\theta^*=\big(c_0^*,\ldots,c^*_{p'^*},0,\ldots,0,d_1^*,\ldots,d^*_{q'^*},0,\ldots,0,a_1^*,\ldots,a^*_{p^*},0,\ldots,0,b_1^*,\ldots,b_{q^*},0,\ldots,0\big)\in\mathbb{R}^d.$$

With Θ a bounded subset of $\Theta_{p_{\max},q_{\max},p'_{\max},q'_{\max}}$, all the previous assumptions D(Θ), Id(Θ), Var(Θ) are satisfied and K(Θ) is also satisfied as soon as $\kappa_n\to\infty$. As a consequence, in this framework the consistency Theorem (3.1) and the Theorem (3.2) of the estimator of the chosen model are satisfied when $r=4$ and $\kappa_n\to\infty$ (for instance with the typical BIC penalty $\kappa_n=\log n$).

5. Portmanteau test

From the above section, we are now able to asymptotically pick up a best model in a family of models. We can also obtain asymptotic confident regions of the estimated parameter of the chosen model. However, it is also important to check whether the chosen model is appropriate. This section attempts to answer this question by constructing a portmanteau test as a diagnostic tool based on the squares of the residuals sequence of the chosen model.

This test has been widely considered in the time series literature, with procedures based on the squared residual correlogram (see for instance [31], [32]) and the absolute residual (or usual residuals) correlogram (see for instance [30], [14], [29]), among others.

Since our goal is to provide an efficient test for the entire affine class that contains weak white noise processes, we consider in this setting the autocorrelation of the squared residuals and then we will follow the same scheme of procedure used in ([31], [32]) while relying on some of their results.

For $m\in\mathcal{M}$, for K a positive integer, denote the vector of adjusted correlogram of squares residuals by:

$$\widehat{\rho}(m):=\big(\widehat{\rho}_1(m),\ldots,\widehat{\rho}_K(m)\big)',$$

where for $k=1,\ldots,K$, $\widehat{\rho}_k(m):=\dfrac{\widehat{\gamma}_k(m)}{\widehat{\gamma}_0(m)}$ with

$$\widehat{\gamma}_k(m):=\frac{1}{n}\sum_{t=k+1}^{n}\big(\widehat{e}_t^{\,2}(m)-1\big)\big(\widehat{e}^{\,2}_{t-k}(m)-1\big)\quad\text{and}\quad \widehat{e}_t(m):=\big(\widehat{M}^t_{\widehat{\theta}(m)}\big)^{-1}\big(X_t-\widehat{f}^t_{\widehat{\theta}(m)}\big).$$

Finally, the following theorem provides central limit theorems for $\widehat{\rho}(m^*)$ and $\widehat{\rho}(\widehat{m})$ as well as for a portmanteau test statistic.

Theorem 5.1. *Under the assumptions of Theorem 3.2, if*

- $\mathbb{E}[\xi_0^3] = 0$;
- $\displaystyle\sum_{t=1}^{\infty} t^{-1/4}\Big(\sum_{j\geq t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)\Big)^{1/2} < \infty$ *or* $\displaystyle\sum_{t=1}^{\infty} t^{-1/4}\Big(\sum_{j\geq t}\alpha_j(\widetilde{H}_\theta,\Theta)\Big)^{1/2} < \infty$;

then,

1. *With $V(\theta^*,m^*)$ defined in (7.37), it holds that*
$$\sqrt{n}\,\widehat{\rho}(m^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\big(0\,,\,V(\theta^*,m^*)\big). \tag{5.1}$$
2. *With $\widehat{Q}_K(m^*) := n\,\widehat{\rho}(m^*)'\big(V(\widehat{\theta}(m^*),m^*)\big)^{-1}\widehat{\rho}(m^*)$, we have*
$$\widehat{Q}_K(m^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K). \tag{5.2}$$
3. *The previous points 1. and 2. also hold when m^* is replaced by $\widehat{m}$.*

Using the Theorem 5.1, we can asymptotically test:

$$\begin{cases} H_0:\ \exists m^*\in\mathcal{M}, \text{ such as } (X_1,\dots,X_n) \text{ is a trajectory of } X\in\mathcal{AC}(M_\theta,f_{\theta^*}) \text{ with } \theta^*\in\Theta(m^*) \\ \\ H_1:\ \nexists m^*\in\mathcal{M}, \text{ such as } (X_1,\dots,X_n) \text{ is a trajectory of } X\in\mathcal{AC}(M_\theta,f_{\theta^*}) \text{ with } \theta^*\in\Theta(m^*) \end{cases}.$$

Therefore, $\widehat{Q}_K(\widehat{m})$ can be used as a portmanteau test statistic to decide between H_0 and H_1 and diagnose the goodness-of-fit of the selected model.

Remark 3.
1. Like in [31], it is important to point out that for $ARCH(p)$ model, since $f_\theta^t = 0$, we have $\mathbb{E}\Big[\big(\xi_0^2-1\big)\,\partial_\theta\log\big(M_{\theta^*}^k\big)\Big] = 0$ for all $k > p$. Hence, for these models, the matrix $V(\theta^*,m^*)-I_K$ will have approximately zero entries from the $(p+1)^{th}$ row onwards and then the standard errors of $\widehat{\rho}(m^*)_i$ are in this case equal to $1/\sqrt{n}$ for $i = p+1,\dots,K$. The statistic $\widehat{Q}_K(m^*)$ yields to $\widehat{Q}(p,K) := n\sum_{i=p+1}^K[\widehat{\rho}(m^*)_i]^2$ which will be asymptotically χ^2 distributed with $K-p$ degrees of freedom.
2. In practice the constant μ_4 and the rows of the matrix $V(\widehat{\theta}(m^*),m^*)$ involved in the previous theorem are estimated by the correspondent sample average; they are respectively $\widehat{\mu}_4 = \frac{1}{n}\sum_{t=1}^n(\widehat{e}_t(\widehat{m}))^4$ and $\big(\widehat{V}(\widehat{\theta}((\widehat{m})),(\widehat{m}))\big)_{k,.} = \frac{1}{n}\sum_{t=k+1}^n[(\widehat{e}_t(\widehat{m}))^2-1][\partial_\theta\log\big(M_\theta^k\big)]_{(\theta=\widehat{\theta}(\widehat{m}))}$.

6. Numerical Results

This section features some simulation experiments that are performed to assess the usefulness of the asymptotic results obtained in Section 3. The various configurations studied are presented below and we compare the performance of penalties $\log n$ and $\sqrt{n}$. The process used to generate the trajectory is indicated each time.

Each model is generated independently 1000 times over a trajectory of length n. Different sample sizes are considered to identify possible discrepancies between asymptotically expected properties and those obtained at finite distance. We will consider n belongs to $\{100, 500, 1000, 2000\}$. Throughout this section, (ξ_t) represents a Gaussian white noise with variance unity.

6.1. Classical configurations

We first simulate some classical model illustrated as follows and the results are displayed in the Table 1.

1. Model 1, AR(2) process: $X_t = 0.4X_{t-1} + 0.4X_{t-2} + \xi_t$.

2. Model 2, ARMA$(1,1)$ process: $X_t = 0.3X_{t-1} + \xi_t + 0.5\xi_{t-1}$.
3. Model 3, ARCH(2) process: $X_t = \xi_t\sqrt{0.2 + 0.4X_{t-1}^2 + 0.2X_{t-2}^2}$.

We considered as competitive models all the models in the family $\mathcal{M}$ defined by:

$$\mathcal{M} = \big\{\text{ARMA}(p,q) \text{ or GARCH}(p',q') \text{ processes with } 0 \le p,q,p' \le 5,\ 1 \le q' \le 5\big\}.$$

As a consequence, there are 66 candidate models.

The Table 1 shows for each penalty ($\log n$ and $\sqrt{n}$) the percentage of times the associated criterion selects respectively a wrong model, the true model and an overfitted model (here a model which contains the true model).

Table 1: Percentage of selected order based on 1000 replications depending on sample's length for Model 1, 2 and 3 respectively.

	Sample length n	100		500		1000		2000	
	Penalty	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$
Model 1	Wrong	21	32.3	3	0.9	0.9	0	0.2	0
	True	74.6	67.5	95.8	99.1	98.2	100	99	100
	Overfitted	4.4	0.2	1.2	0	0.9	0	0.8	0
Model 2	Wrong	81.8	97.5	30.1	67.4	19.9	33.2	10.2	10.5
	True	16.1	2.5	69.1	32.6	79.5	66.8	89.4	89.5
	Overfitted	2.1	0	0.8	0	0.6	0	0.4	0
Model 3	Wrong	78.9	92.9	25.7	70.5	11.6.	39.2	5.4	11.4
	True	20.4	7.0	73.2	29.5	88.1	60.8	94.3	88.6
	Overfitted	0.1	0.1	1.1	0	0.3	0	0.3	0

From these results, it is clear that the consistency of our model selection procedure is numerically convincing, which is in accordance with Theorem 3.1, where both the criteria are consistent for Model 1, 2 and 3. Note also that the typical BIC $\log n$ penalty is the most interesting for retrieving the true model than the $\sqrt{n}$-penalized likelihood for a small sample size. But the larger the sample size, the more accurate the $\sqrt{n}$ penalty case.

For each of the three models, we also applied the portmanteau test statistic $\widehat{Q}_K(\widehat{m})$, using the $\sqrt{n}$ penalty. Table 2 shows the empirical size and empirical power of this test. We call by empirical size, the percentage of falsely rejecting the null hypothesis H_0. On the other hand, the empirical power represents the percentage of rejection of H_0 when we arbitrary chose a false model, which is a AR(3) process $X_t = 0.2X_{t-1} + 0.2X_{t-2} + 0.4X_{t-1} + \xi_t$ for Model 1 and 2, and a ARCH(3) process $X_t = \xi_t\sqrt{0.4 + 0.2X_{t-1}^2 + 0.2X_{t-2}^2 + 0.2X_{t-3}^2}$ for Model 3.

It is important to note that choosing the maximum number of lags K is sometimes tricky. To our knowledge, there is no real theoretical study to justify the choice of one value or another. However, some Monte Carlo simulations have suggested some ways to make a good choice . For instance [31] suggested that the autocorrelations $\widehat{\rho}_k(\widehat{m})$ with $1 \le k \le K$ have a better asymptotic behaviour for small values of k. Therefore, the finite sample performance of the size and power of the test may also vary with the choice of K and could be better for small values of K. On the other hand, [47] suggested that $K = p + q + 1$ may be an appropriate choice for the GARCH(p,q) family.
Thus, in our tests, we consider $K = 3$ and $K = 6$ so that the rejection is based on the upper 5th percentile of the $\chi^2(3)$ distribution on the one hand and $\chi^2(6)$ on the other hand.

Table 2: The empirical size and empirical power of the portmanteau test statistic $\widehat{Q}_K(\widehat{m})$ based on 1000 independent replications (in %) with $K = 3$ and $K = 6$.

Sample length		100		500		1000		2000	
		size	power	size	power	size	power	size	power
$K = 3$	Model 1	3.5	13.6	3.8	48.1	3.5	82.7	3.2	97.7
	Model 2	4.0	6.7	5	21.7	4.8	38.6	4.4	64.2
	Model 3	4.3	52.7	4.2	98.6	3.2	99.6	3.6	99.9
$K = 6$	Model 1	3.5	9.4	4.8	43.1	5.3	74.6	4.5	97.6
	Model 2	2.1	6.3	4.9	18	4.5	32.2	6.4	61.3
	Model 3	3	18.3	3.1	91.5	3.4	99.6	6.8	99.7

Once again, the results of Table 2 numerically confirms the asymptotic results of Theorem 5.1. Remark that the test is more powerful by using values of K not too large as mentioned above especially for small samples.

6.2. *Subset model selection*

Now, we exhibit the performance of the criteria on a particular case of dimension selection. The process generated data is considered as follows:

$$\text{Model 4} : X_t = 0.4 X_{t-3} + 0.4 X_{t-4} + \xi_t .$$

Here, we will consider the case of a nonhierarchical but exhaustive family $\mathcal{M}$ of AR(4) models , *i.e.*

$$\begin{aligned} \mathcal{M} &= \mathcal{P}(\{1,2,3,4\}) \\ &\Longrightarrow \quad X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \theta_3 X_{t-3} + \theta_4 X_{t-4} + \xi_t \text{ and } \theta = (\theta_1, \theta_2, \theta_3, \theta_4)' \in \Theta(m). \end{aligned}$$

As a consequence, $16 = 2^4$ candidate models are considered and Table 3 presents the results of the selection procedure.

Table 3: Percentage of selected model based on 1000 replications depending on sample's length for Model 4

Sample length	100		500		1000		2000	
	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$
true model	85.9	68	97.5	100	96.8	100	98.9	100
overfitted	7.9	2	2.5	0	3.2	0	1.1	0
false model	6.2	30	0	0	0	0	0	0

We deduce that the consistency of our model selection procedure is also numerically convincing in this case of exhaustive model selection, which is in accordance with Theorem 3.1

6.3. *Slow decrease of the Lipschitz coefficients*

In this subsection, we consider an $AR(2) - ARCH(\infty)$ with a slow decrease of its Lipschitz coefficients in order to numerically show that the penalty $\log n$ is not consistent in all cases. The considered data generating process is featured as follows:

$$\text{Model 5} : X_t = -0.45\, X_{t-1} + 0.4\, X_{t-2} + \xi_t \text{ with } \xi_t = \varepsilon_t \sqrt{0.5 + 0.1 \sum_{i \geq 1} \xi_{t-i}^2 / i^3},$$

where ε_t is an i.i.d random sequence with mean 0 and variance 1. The sequence $(\alpha_i)_{i \geq 1}$ verifies $\alpha_i = O(i^{-3})$ so that the sequence of Lipschitz coefficients of M_θ^ξ is given by $\alpha_i(M_\theta^\xi) = O(i^{-1.5})$ and then the decrease rate of the sequence $\big(\alpha_i(M_\theta^X)\big)$ is equal to $O(i^{-1.5})$. From Remark 1, all penalties such as n^δ with $\delta > 2 - 1.5 = 0.5$ will lead to a consistent model selection criterion and this is not the case for the typical BIC $\log n$ penalty. We have considered $\delta = 2/3$ as in the Bridge Criteria (BC) recently proposed in [10]. Here the family of model $\mathcal{M}$ is defined by

$$\mathcal{M} = \big\{ \text{AR}(p)\text{-ARCH}(\infty) \text{ processes with } 1 \leq p \leq 8, \text{ where the ARCH}(\infty) \text{ is defined as in Model 5} \big\}.$$

The results of simulations are featured in Table 4.

Table 4: Percentage of selected order based on 1000 replications depending on sample's length for model 5

Sample length	100		500		1000		2000	
	$\log n$	$n^{2/3}$	$\log n$	$n^{2/3}$	$\log n$	$n^{2/3}$	$\log n$	$n^{2/3}$
$p<2$	8.9	69.1	0.1	17.6	0	6.1	0	0.9
$p=2$	88.9	30.9	75.3	82.4	71.4	93.9	72.5	99.1
$p>2$	2.2	0	24.6	0	28.6	0	27.5	0

Note that we also computed the $\log n$ criterion for $n = 5000$ and $n = 10000$ in additional numerical experiments and its frequencies of choice of the true order $p = 2$ were almost 72%. As a consequence, for this model selection framework of the infinite memory process with a slow decrease of Lipschitz coefficients, the usual BIC penalty $\log n$ seems numerically not sufficient to avoid overfitting in contrast with a $n^{2/3}$ penalty that leads to a consistent criterion.

6.4. Illustrative Example

We consider the returns of the daily closing prices of the FTSE index of the London Stock Exchange 100. They are 2273 observations from January 4th, 2010 to December 31st, 2018. The mean and standard deviation of the returns are -0.54 and 57.67, respectively. The Time plot and the correlograns for the log returns and squared log returns are plotted in Figure 1.

The Figures (1a) and (1c) exhibit the conditional heteroskedasticity in the log return time series. Moreover, Figure (1b) shows that more than 5 per cent of the autocorrelations are out of the confidence interval $\pm 1.96/\sqrt{2273}$ and specially the Figure (1d) suggests that the strong white noise assumption cannot be sustained for this log-returns sequence of FTSE index.

Therefore, the $GARCH(p,q)$ family was considered for the modelling of the FTSE index with $(p,q) \in [\![1;10]\!] \times [\![0;10]\!]$ which lead us to 110 candidate models. The penalization $\log n$ and $\sqrt{n}$ have been applied to identify the best order and the goodness-of-fit of the selected model has been tested by the portmanteau test. Based on the results of the simulations, we set $K = 3$ for the portmanteau test statistic.

The GARCH$(1,1)$ is the "best" model according to both criteria (related to above penalizations) and the portmanteau statistic $\widehat{Q}_3(\widehat{m}) \simeq 2.13$ is associated with a p-value of 0.55. Hence, the selected model GARCH$(1,1)$ is adequate to model the FTSE 100 index using either criterion.

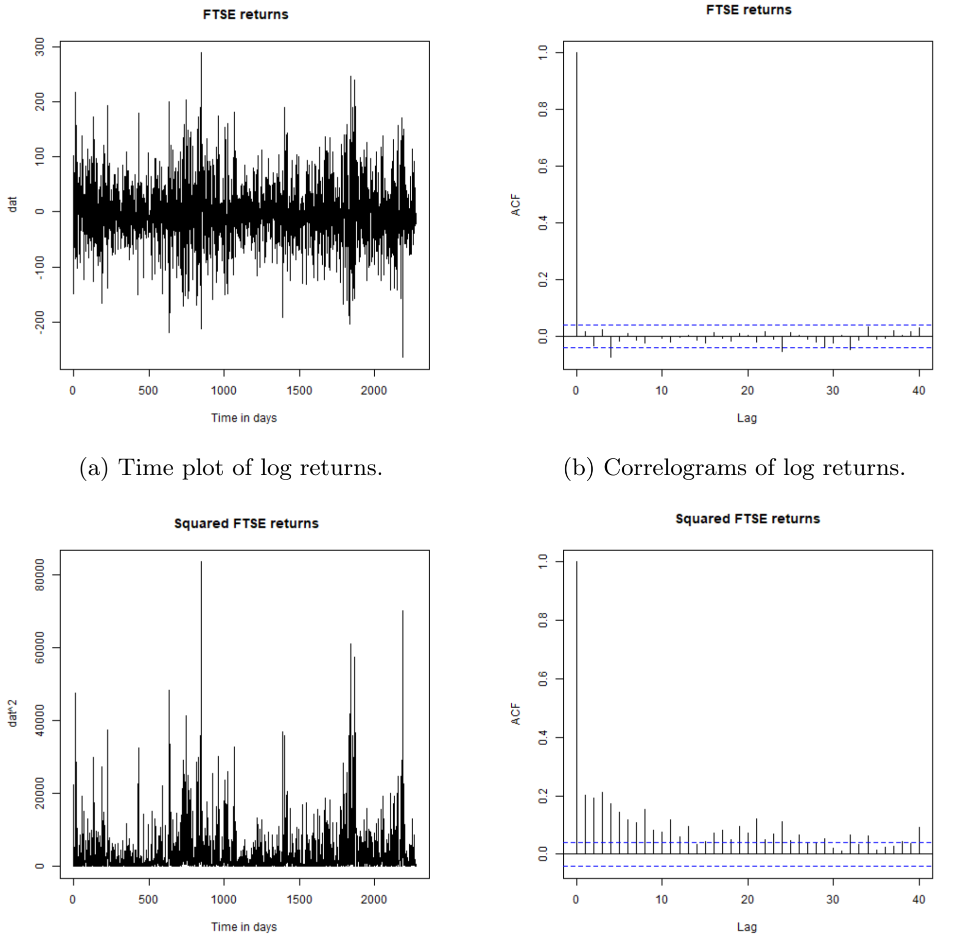
7. Proofs

We start with the proof of the Proposition 1.

Proof. For ease of writing, consider only the general case where $f^{(i)}_{\theta_i} = g^{(i)}_{\alpha_i}$ and $M^{(i)}_{\theta_i} = N^{(i)}_{\beta_i}$ where $\theta_i = {}^t(\alpha_i, \beta_i)$ for $i = 1, 2$. Now, assume that there exist $\alpha \in \mathbb{R}^\delta$, where $0 \le \delta \le \min(d_1, d_2)$ and a function h_α such as $g^{(1)}_{\alpha_1} = h_\alpha + \ell^{(1)}_{\alpha'_1}$, $f^{(2)}_{\alpha_2} = h_\alpha + \ell^{(2)}_{\alpha'_2}$ with $\alpha_1 = {}^t(\alpha, \alpha'_1)$ and $\alpha_2 = {}^t(\alpha, \alpha'_2)$ and $\ell^{(i)}_0 = 0$. Similarly, assume that there exist $\beta \in \mathbb{R}^{\delta'}$, where $0 \le \delta' \le \min(d_1, d_2)$ and a function R_β such as $N^{(1)}_{\beta_1} = R_\beta + m^{(1)}_{\beta'_1}$, $N^{(2)}_{\beta_2} = R_\beta + m^{(2)}_{\beta'_2}$ with $\beta_1 = {}^t(\beta, \beta'_1)$ and $\beta_2 = {}^t(\beta, \beta'_2)$ and $m^{(i)}_0 = 0$.
Consider now $\theta = {}^t(\alpha, \alpha'_1, \alpha'_2, \beta, \beta'_1, \beta'_2) \in \mathbb{R}^d$ (and therefore $\max(d_1, d_2) \le d \le d_1 + d_2$), $f_\theta = h_\alpha + \ell^{(1)}_{\alpha'_1} + \ell^{(2)}_{\alpha'_2}$ and $M_\theta = R_\beta + m^{(1)}_{\beta'_1} + m^{(2)}_{\beta'_2}$. Then if $X \in \mathcal{AC}\big(M_\theta, f_\theta\big)$, for any $t \in \mathbb{Z}$,

$$X_t = \big(R_\beta((X_{t-k})_{k\ge1}) + m^{(1)}_{\beta'_1}((X_{t-k})_{k\ge1}) + m^{(2)}_{\beta'_2}((X_{t-k})_{k\ge1})\big)\,\xi_t \\ + \big(h_\alpha((X_{t-k})_{k\ge1}) + \ell^{(1)}_{\alpha'_1}((X_{t-k})_{k\ge1}) + \ell^{(2)}_{\alpha'_2}((X_{t-k})_{k\ge1})\big).$$

Then, for $\alpha'_2 = \beta'_2 = 0$, $X \in \mathcal{AC}\big(M^{(1)}_{\theta_1}, f^{(1)}_{\theta_1}\big)$ and for $\alpha'_1 = \beta'_1 = 0$, $X \in \mathcal{AC}\big(M^{(2)}_{\theta_2}, f^{(2)}_{\theta_2}\big)$. ∎



(a) Time plot of log returns.

(b) Correlograms of log returns.

(c) Time plot of squared log returns.

(d) Correlograms of squared log returns.

Figure 1: Daily closing FTSE 100 index (January 4th, 2010 to December 31 st, 2018).

In the sequel, some lemmas are stated and theirs proofs are given.

Lemma 1. *Let $X \in \mathcal{AC}(M_\theta, f_\theta)$ (or $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$) and $\Theta \subseteq \Theta(r)$ (or $\Theta \subseteq \widetilde{\Theta}(r)$) with $r \geq 2$. Assume that the assumptions $D(\Theta)$ and $K(\Theta)$ (or $\widetilde{K}(\Theta)$) hold. Then:*

$$\frac{1}{\kappa_n} \big\| \widehat{L}_n(\theta) - L_n(\theta) \big\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0. \tag{7.1}$$

Proof. We have $|\widehat{L}_n(\theta) - L_n(\theta)| \leq \sum_{t=1}^n |\widehat{q}_t(\theta) - q_t(\theta)|$. Then,

$$\frac{1}{\kappa_n} \big\| \widehat{L}_n(\theta) - L_n(\theta) \big\|_\Theta \leq \frac{1}{\kappa_n} \sum_{t=1}^n \|\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta.$$

By Corollary 1 of [27], with $r \leq 3$, (7.1) is established when:

$$\sum_{k \geq 1} (\frac{1}{\kappa_k})^{r/3} \mathbb{E}\big(\|\widehat{q}_k(\theta) - q_k(\theta)\|_\Theta^{r/3} \big) < \infty. \tag{7.2}$$

With $r \geq 3$, and under the assumptions, we first recall some results already obtained in [7]: for any $t \in \mathbb{Z}$,

- $$\mathbb{E}\big[|X_t|^r + \|f_\theta^t\|_\Theta^r + \|\widehat{f}_\theta^t\|_\Theta^r + \|M_\theta^t\|_\Theta^r + \|\widehat{M}_\theta^t\|_\Theta^r + \|H_\theta^t\|_\Theta^{r/2} + \|\widehat{H}_\theta^t\|_\Theta^{r/2} \big] < \infty \tag{7.3}$$
- $$\begin{cases} \mathbb{E}\big[\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^r \big] \leq C \left(\sum_{j \geq t} \alpha_j(f_\theta, \Theta) \right)^r \\ \mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r \big] \leq C \left(\sum_{j \geq t} \alpha_j(M_\theta, \Theta) \right)^r \\ \mathbb{E}\big[\|H_\theta^t - \widehat{H}_\theta^t\|_\Theta^{r/2} \big] \leq C \left(\min \left\{ \sum_{j \geq t} \alpha_j(M_\theta, \Theta) \ , \ \sum_{j \geq t} \alpha_j(H_\theta, \Theta) \right\} \right)^{r/2}. \end{cases} \tag{7.4}$$

For any $\theta \in \Theta$, we have:

$$\begin{aligned} &|\widehat{q}_t(\theta) - q_t(\theta)| = \Big| \frac{(X_t - \widehat{f}_\theta^t)^2}{\widehat{H}_\theta^t} + \log(\widehat{H}_\theta^t) - \frac{(X_t - f_\theta^t)^2}{H_\theta^t} - \log(H_\theta^t) \Big| \\ &\leq (H_\theta^t \widehat{H}_\theta^t)^{-1} \big| H_\theta^t (X_t - \widehat{f}_\theta^t)^2 - \widehat{H}_\theta^t (X_t - f_\theta^t)^2 \big| + \big| \log(\widehat{H}_\theta^t) - \log(H_\theta^t) \big| \\ &\leq (H_\theta^t \widehat{H}_\theta^t)^{-1} \big| (H_\theta^t - \widehat{H}_\theta^t)(X_t - f_\theta^t)^2 - H_\theta^t (X_t - f_\theta^t)^2 + H_\theta^t (X_t - \widehat{f}_\theta^t)^2 \big| + \big| \log(\widehat{H}_\theta^t) - \log(H_\theta^t) \big| \\ &\leq \underline{h}^{-3/2} \big(|X_t|^2 + 2|X_t| |f_\theta^t| + |f_\theta^t|^2 \big) \big| M_\theta^t - \widehat{M}_\theta^t \big| + \underline{h}^{-1} \big(2|X_t| + |f_\theta^t| + |\widehat{f}_\theta^t| \big) \big| f_\theta^t - \widehat{f}_\theta^t \big| + 2 \big| \log(\widehat{M}_\theta^t) - \log(M_\theta^t) \big| \\ &\leq \underline{h}^{-3/2} \big(|X_t|^2 + 2|X_t| \times \|f_\theta^t\|_\Theta + \|f_\theta^t\|_\Theta^2 \big) \|M_\theta^t - \widehat{M}_\theta^t\|_\Theta \\ &\qquad\qquad + \underline{h}^{-1} \big(2|X_t| + \|f_\theta^t\|_\Theta + \|\widehat{f}_\theta^t\|_\Theta \big) \|f_\theta^t - \widehat{f}_\theta^t\|_\Theta + 2\, \underline{h}^{-1/2} \|\widehat{M}_\theta^t - M_\theta^t\|_\Theta. \end{aligned}$$

1/ If $X \subset \mathcal{AC}(M_\theta, f_\theta)$, we deduce

$$\begin{aligned} \mathbb{E}\big[\|\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta^{r/3} \big] \leq C \Big(&\mathbb{E}\Big[\big(\|X_t + f_\theta^t\|_\Theta^2 + 1 \big)^{r/3} \|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^{r/3} \Big] \\ &+ \mathbb{E}\Big[\big(2|X_t| + \|f_\theta^t\|_\Theta + \|\widehat{f}_\theta^t\|_\Theta \big)^{r/3} \|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^{r/3} \Big] \Big). \end{aligned} \tag{7.5}$$

Then, by Hölder's inequality and (7.3) we have:

$$\begin{aligned} &\mathbb{E}\Big[\big(\|X_t + f_\theta^t\|_\Theta^2 + 1 \big)^{r/3} \|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^{r/3} \Big] \\ &\qquad \leq \Big(\mathbb{E}\big[\|X_t + f_\theta^t + 1\|_\Theta^r \big] \Big)^{2/3} \Big(\mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r \big] \Big)^{1/3} \leq C \, \Big(\mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r \big] \Big)^{1/3}. \end{aligned} \tag{7.6}$$

Again with Hölder's inequality and (7.3) ,

$$\mathbb{E}\big[\big((2|X_t| + \|f_\theta^t\|_\Theta + \|\widehat{f}_\theta^t\|_\Theta) \|f_\theta^t - \widehat{f}_\theta^t\|_\Theta \big)^{r/3} \big] \leq C \, \big(\mathbb{E}\big[\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^r \big] \big)^{1/3}. \tag{7.7}$$

Therefore, from (7.6), (7.7) and (7.4), there exists a constant C such that

$$\mathbb{E}\big[\|(\widehat{q}_t(\theta)-q_t(\theta)\|_\Theta^{r/3}\big] \le C\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\sum_{j\ge t}\alpha_j(M_\theta,\Theta)\Big)^{r/3}. \tag{7.8}$$

Hence,

$$\sum_{k\ge 1}(\frac{1}{\kappa_k})^{r/3}\mathbb{E}\big[\|\widehat{q}_k(\theta)-q_k(\theta)\|_\Theta^{r/3}\big] \le C\sum_{k\ge 1}(\frac{1}{\kappa_k})^{r/3}\Big(\sum_{j\ge k}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)\Big)^{r/3},$$

which is finite by assumption $K(\Theta)$, and this achieves the proof.

2/ If $X\subset\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$ and using Corollary 1 of [27], with $r\le 4$, (7.1) is established when:

$$\sum_{k\ge 1}(\frac{1}{\kappa_k})^{r/4}\mathbb{E}\big(\|\widehat{q}_k(\theta)-q_k(\theta)\|_\Theta^{r/4}\big)<\infty. \tag{7.9}$$

By proceeding as in the previous case, we deduce

$$|\widehat{q}_t(\theta)-q_t(\theta)| \le \underline{h}^{-2}|X_t|^2\,\|H_\theta^t-\widehat{H}_\theta^t\|_\Theta+\underline{h}^{-1}\|\widehat{H}_\theta^t-H_\theta^t\|_\Theta.$$

In addition, we deduce that there exists a constant C such that

$$\mathbb{E}\big[\|(\widehat{q}_t(\theta)-q_t(\theta)\|_\Theta^{r/4}\big] \le C\Big(\sum_{j\ge t}\alpha_j(H_\theta,\Theta)\Big)^{r/4}. \tag{7.10}$$

■

Lemma 2. *Let $X\in\mathcal{AC}(M_\theta,f_\theta)$ (or $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$) and $\Theta\subseteq\Theta(r)$ (or $\Theta\subseteq\widetilde{\Theta}(r)$) with $r\ge 2$. Assume that the assumptions $D(\Theta)$ and $K(\Theta)$ (or $\widetilde{K}(\Theta)$) hold. Then:*

$$\frac{1}{\kappa_n}\Big\|\frac{\partial\widehat{L}_n(\theta)}{\partial\theta}-\frac{\partial L_n(\theta)}{\partial\theta}\Big\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0. \tag{7.11}$$

Proof. We will go along similar lines as in the proof of Lemma 1. We have:

$$\frac{1}{\kappa_n}\Big\|\frac{\partial\widehat{L}_n(\theta)}{\partial\theta}-\frac{\partial L_n(\theta)}{\partial\theta}\Big\|_\Theta \le \frac{1}{\kappa_n}\sum_{t=1}^n\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta.$$

Using again Corollary 1 of [27], it is sufficient to prove for $r\le 3$ that

$$\sum_{k\ge 1}(\frac{1}{\kappa_k})^{r/3}\,\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/3}\Big]<\infty. \tag{7.12}$$

For any $\theta\in\Theta$, with $H_\theta=M_\theta^2$, the first partial derivatives of $q_t(\theta)$ are

$$\begin{aligned}\frac{\partial q_t(\theta)}{\partial\theta_i} &= \frac{-2(X_t-f_\theta^t)}{H_\theta^t}\frac{\partial f_\theta^t}{\partial\theta_i}-\frac{(X_t-f_\theta^t)^2}{(H_\theta^t)^2}\frac{\partial H_\theta^t}{\partial\theta_i}+\frac{1}{H_\theta^t}\frac{\partial H_\theta^t}{\partial\theta_i}\\ &= -2(H_\theta^t)^{-1}(X_t-f_\theta^t)\frac{\partial f_\theta^t}{\partial\theta_i}+(X_t-f_\theta^t)^2\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}+(H_\theta^t)^{-1}\frac{\partial H_\theta^t}{\partial\theta_i},\end{aligned}$$

for $i=1,\cdots,d$. Hence,

$$\begin{aligned}\Big|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big| &\le 2\Big|(h_\theta^t)^{-1}(X_t-f_\theta^t)\frac{\partial f_\theta^t}{\partial\theta_i}-(\widehat{h}_\theta^t)^{-1}(X_t-\widehat{f}_\theta^t)\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big|\\ &\quad+\Big|(X_t-\widehat{f}_\theta^t)^2\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-(X_t-f_\theta^t)^2\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big|+\Big|(\widehat{H}_\theta^t)^{-1}\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-(H_\theta^t)^{-1}\frac{\partial H_\theta^t}{\partial\theta_i}\Big|.\end{aligned}$$

Then, using $|a_1b_1c_1-a_2b_2c_2|\le|a_1-a_2|\,|b_2|\,|c_2|+|a_1|\,|b_1-b_2|\,|c_2|+|a_1|\,|b_1|\,|c_1-c_2|$ for any a_1,a_2,b_1,b_2,c_1,c_2 in $\mathbb{R}$, we obtain

$$\begin{aligned}\Big|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big| \le 2\Big(&\big|(H_\theta^t)^{-1}-(\widehat{H}_\theta^t)^{-1}\big|\times\big|X_t-\widehat{f}_\theta^t\big|\,\Big|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big|+\big|(H_\theta^t)^{-1}\big|\times\big|\widehat{f}_\theta^t-f_\theta^t\big|\,\Big|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big|\\ &+\big|(H_\theta^t)^{-1}\big|\times\big|X_t-f_\theta^t\big|\,\Big|\frac{\partial f_\theta^t}{\partial\theta_i}-\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big|\Big)+\big|X_t-\widehat{f}_\theta^t\big|^2\,\Big|\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big|\\ &+2\,\Big|\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big|\,\big|X_t\big|\,\big|f_\theta^t-\widehat{f}_\theta^t\big|+\big|(\widehat{H}_\theta^t)^{-1}\big|\,\Big|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-\frac{\partial H_\theta^t}{\partial\theta_i}\Big|+\Big|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}\Big|\,\big|(\widehat{H}_\theta^t)^{-1}-(H_\theta^t)^{-1}\big|.\end{aligned}$$

Thus,

$$\begin{aligned}\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta &\le 2\,\underline{h}^{-1}\Big(\big\|\widehat{f}_\theta^t-f_\theta^t\big\|_\Theta\Big\|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta+\big\|X_t-f_\theta^t\big\|_\Theta\Big\|\frac{\partial f_\theta^t}{\partial\theta_i}-\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta\Big)\\ &+2\,\big\|(H_\theta^t)^{-1}-(\widehat{H}_\theta^t)^{-1}\big\|_\Theta\big\|X_t-\widehat{f}_\theta^t\big\|_\Theta\Big\|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta+\big\|X_t-\widehat{f}_\theta^t\big\|^2\Big\|\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|\\ &+2\,\big|X_t\big|\,\big\|f_\theta^t-\widehat{f}_\theta^t\big\|_\Theta\Big\|\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta+\big\|(\widehat{H}_\theta^t)^{-1}\big\|_\Theta\Big\|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta+\big\|(\widehat{H}_\theta^t)^{-1}-(H_\theta^t)^{-1}\big\|_\Theta\Big\|\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta.\end{aligned}$$

Using again the results of [7], we know that:

- $$\mathbb{E}\Big[\Big\|\frac{\partial f_\theta^t}{\partial\theta_i}\Big\|_\Theta^r+\Big\|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta^r+\Big\|\frac{\partial M_\theta^t}{\partial\theta_i}\Big\|_\Theta^r+\Big\|\frac{\partial\widehat{M}_\theta^t}{\partial\theta_i}\Big\|_\Theta^r+\Big\|\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/2}+\Big\|\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta^r\Big]<\infty \qquad (7.13)$$
- $$\begin{cases}\mathbb{E}\big[\big\|(H_\theta^t)^{-1}-(\widehat{H}_\theta^t)^{-1}\big\|_\Theta^r\big]\le C\Big(\displaystyle\sum_{j\ge t}\alpha_j(M_\theta,\Theta)\Big)^r\\ \mathbb{E}\Big[\Big\|\dfrac{\partial f_\theta^t}{\partial\theta_i}-\dfrac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta^r\Big]\le C\Big(\displaystyle\sum_{j\ge t}\alpha_j(\partial f_\theta,\Theta)\Big)^r\\ \mathbb{E}\Big[\Big\|\dfrac{\partial H_\theta^t}{\partial\theta_i}-\dfrac{\partial\widehat{H}_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\le C\Big(\displaystyle\sum_{j\ge t}\big(\alpha_j(M_\theta,\Theta)+\alpha_j(\partial M_\theta,\Theta)\big)\Big)^{r/2}\\ \mathbb{E}\Big[\Big\|\dfrac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}-\dfrac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\le C\Big(\displaystyle\sum_{j\ge t}\big(\alpha_j(M_\theta,\Theta)+\alpha_j(\partial M_\theta,\Theta)\big)\Big)^{r/2}\end{cases} \qquad (7.14)$$

1. If $X\subset\mathcal{AC}(M_\theta,f_\theta)$, we deduce from the Hölder's Inequality that,

$$\begin{aligned}\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/3}\Big] &\le C\,\Big[\big(\mathbb{E}\big[\big\|\widehat{f}_\theta^t-f_\theta^t\big\|_\Theta^r\big]\big)^{1/3}\Big(\mathbb{E}\Big[\Big\|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\Big)^{2/3}\\ &+\big(\mathbb{E}\big[\big\|X_t-f_\theta^t\big\|_\Theta^{2r/3}\big]\big)^{1/2}\Big(\mathbb{E}\Big[\Big\|\frac{\partial f_\theta^t}{\partial\theta_i}-\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta^r\Big]\Big)^{1/3}\\ &+\big(\mathbb{E}\big[\big\|(H_\theta^t)^{-1}-(\widehat{H}_\theta^t)^{-1}\big\|_\Theta^r\big]\big)^{1/3}\Big(\mathbb{E}\big[\big\|X_t-\widehat{f}_\theta^t\big\|_\Theta^r\big]\,\mathbb{E}\Big[\Big\|\frac{\partial\widehat{f}_\theta^t}{\partial\theta_i}\Big\|_\Theta^r\Big]\Big)^{1/3}\\ &+\big(\mathbb{E}\big[\big\|X_t-\widehat{f}_\theta^t\big\|_\Theta^r\big]\big)^{1/3}\Big(\mathbb{E}\Big[\Big\|\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|^{r/2}\Big]\Big)^{2/3}\\ &+\Big(\mathbb{E}\Big[\Big\|\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta^r\Big]\Big)^{1/3}\Big(\mathbb{E}\big[|X_t|^r\big]\,\mathbb{E}\big[\big\|f_\theta^t-\widehat{f}_\theta^t\big\|_\Theta^r\big]\Big)^{1/3}\\ &+\Big(\mathbb{E}\Big[\Big\|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/3}\Big]+\Big(\mathbb{E}\Big[\Big\|\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\Big)^{2/3}\big(\mathbb{E}\big[\big\|(\widehat{H}_\theta^t)^{-1}-(H_\theta^t)^{-1}\big\|_\Theta^r\big]\Big)^{1/3}\Big].\end{aligned}$$

Using (7.13) and (7.14), we deduce

$$\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/3}\Big]\ \le\ C\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)\ +\ \alpha_j(M_\theta,\Theta)\ +\ \alpha_j(\partial f_\theta,\Theta)\ +\ \alpha_j(\partial M_\theta,\Theta)\Big)^{r/3}.$$

Therefore,

$$\sum_{k\geq 1}\frac{1}{\kappa_k^{r/3}}\,\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_k(\theta)}{\partial\theta_i}-\frac{\partial q_k(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/3}\Big] \leq C\sum_{k\geq 1}\frac{1}{\kappa_k^{r/3}}\Big(\sum_{j\geq t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\partial f_\theta,\Theta)+\alpha_j(\partial M_\theta,\Theta)\Big)^{r/3}.$$

We conclude the proof of (7.12) from assumption $K(\Theta)$.

2. If $X\subset\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$, we deduce

$$\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta\leq\big|X_t\big|^2\,\Big\|\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta + \underline{h}^{-1}\Big\|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta+\big\|(\widehat{H}_\theta^t)^{-1}-(H_\theta^t)^{-1}\|_\Theta\Big\|\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta.$$

As a consequence,

$$\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/4}\Big]\leq\Big(\mathbb{E}\big[\big|X_t\big|^r\,\mathbb{E}\Big[\Big\|\frac{\partial(\widehat{H}_\theta^t)^{-1}}{\partial\theta_i}-\frac{\partial(H_\theta^t)^{-1}}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\Big)^{1/2} + \underline{h}^{-r/4}\mathbb{E}\Big[\Big\|\frac{\partial\widehat{H}_\theta^t}{\partial\theta_i}-\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/4}\Big]+\Big(\mathbb{E}\big[\|(\widehat{H}_\theta^t)^{-1}-(H_\theta^t)^{-1}\|_\Theta^{r/2}\big]\,\mathbb{E}\Big[\Big\|\frac{\partial H_\theta^t}{\partial\theta_i}\Big\|_\Theta^{r/2}\Big]\Big)^{1/2},$$

implying

$$\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i}-\frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_\Theta^{r/4}\Big]\leq C\Big(\sum_{j\geq t}\alpha_j(H_\theta,\Theta)+\alpha_j(\partial H_\theta,\Theta)\Big)^{r/4},$$

which achieves the proof, according to Corollary 1 of [27]. ■

Lemma 3. *Under the assumptions of Theorem 3.1 and if a model $m\in\mathcal{M}$ is such that $\theta^*\in\Theta(m)$, then:*

$$\frac{1}{\kappa_n}\big|\widehat{L}_n(\widehat{\theta}(m))-\widehat{L}_n(\theta^*)\big|=o_P(1). \tag{7.15}$$

Proof. We have:

$$\frac{1}{\kappa_n}\big|\widehat{L}_n(\widehat{\theta}(m))-\widehat{L}_n(\theta^*)\big|=\frac{1}{\kappa_n}\big|\widehat{L}_n(\widehat{\theta}(m))-L_n(\widehat{\theta}(m))+L_n(\widehat{\theta}(m))-L_n(\theta^*)+L_n(\theta^*)-\widehat{L}_n(\theta^*)\big| \leq\frac{2}{\kappa_n}\big\|\widehat{L}_n(\theta)-L_n(\theta)\big\|_{\Theta(r)}+\frac{1}{\kappa_n}\big|L_n(\widehat{\theta}(m))-L_n(\theta^*)\big|.$$

According to Lemma 1, $\frac{1}{\kappa_n}\big\|\widehat{L}_n(\theta)-L_n(\theta)\big\|_{\Theta(r)}\xrightarrow[n\to+\infty]{a.s.}0$. The proof will be achieved if we prove

$$\frac{1}{\kappa_n}\big|L_n(\widehat{\theta}(m))-L_n(\theta^*)\big|=o_P(1). \tag{7.16}$$

Applying a second order Taylor expansion of L_n around $\widehat{\theta}(m)$ for n sufficiently large such that $\overline{\theta}(m)\in\Theta(m)$ which are between $\widehat{\theta}(m)$ and θ^*, yields:

$$\frac{1}{\kappa_n}\big(L_n(\widehat{\theta}(m))-L_n(\theta^*)\big)= \frac{1}{\kappa_n}\big(\widehat{\theta}(m)-\theta^*\big)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}+\frac{1}{2\kappa_n}\big(\widehat{\theta}(m)-\theta^*\big)'\,\frac{\partial^2L_n(\overline{\theta}(m))}{\partial\theta^2}\,\big(\widehat{\theta}(m)-\theta^*\big). \tag{7.17}$$

Let us deal first with the first term on the right hand side of last equality:

$$\frac{1}{\kappa_n}\left(\widehat{\theta}(m)-\theta^*\right)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}=\frac{1}{\kappa_n}\sqrt{n}\big(\widehat{\theta}(m)-\theta^*\big)\,\frac{1}{\sqrt{n}}\,\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}.$$

Since $\frac{1}{\kappa_n}=o(1)$ and from [7] we have $\sqrt{n}\big(\widehat{\theta}(m)-\theta^*\big)=O_P(1)$ and $\frac{1}{\sqrt{n}}\,\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}=o_P(1)$, it follows that:

$$\frac{1}{\kappa_n}\left(\widehat{\theta}(m)-\theta^*\right)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}=o_P(1). \tag{7.18}$$

On the other hand, for the second term of the right hand side of equality (7.17), let us note that, we have from [7]:

- $\sqrt{n}\left(\widehat{\theta}(m)-\theta^*\right)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{A}_{\theta^*,m}$ a Gaussian random variable from (3.3).
- $-\frac{2}{n}\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\xrightarrow[n\to+\infty]{a.s.}F(\theta^*,m)$ since $\widehat{\theta}(m)\xrightarrow[n\to+\infty]{a.s.}\theta^*$ and using the assumption Var(Θ) insuring that the matrix $F(\theta^*,m)$ exists and is definite positive (see [7]).

Hence,

$$\begin{aligned}&\big(\widehat{\theta}(m)-\theta^*\big)'\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\big(\widehat{\theta}(m)-\theta^*\big)\\&\qquad=\frac{-1}{2}\sqrt{n}\big(\widehat{\theta}(m)-\theta^*\big)'\left(F(\theta^*,m)+o_P(1)\right)\sqrt{n}\big(\widehat{\theta}(m)-\theta^*\big)\\&\qquad\qquad\xrightarrow[n\to\infty]{\mathcal{P}}\frac{-1}{2}\mathcal{A}'_{\theta^*,m}\,F(\theta^*,m)\,\mathcal{A}_{\theta^*,m}.\end{aligned}$$

We deduce that

$$\begin{aligned}&\big(\widehat{\theta}(m)-\theta^*\big)'\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\big(\widehat{\theta}(m)-\theta^*\big)=O_P(1)\\&\qquad\Longrightarrow\ \frac{1}{\kappa_n}\big(\widehat{\theta}(m)-\theta^*\big)'\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\big(\widehat{\theta}(m)-\theta^*\big)=o_P(1).\end{aligned} \tag{7.19}$$

Thus, (7.16) follows from (7.17), (7.18) and (7.19); which completes the proof of Lemma 3. ■

7.1. Misspecified model

When a model m is misspecified, we will show that $\mathbb{P}(\widehat{m}=m^*)\xrightarrow[n\to\infty]{}0$ following the same scheme of proof than in [42]. Before dealing with this proof, we state some useful results.

Proposition 2. *Let $X\in\mathcal{AC}(M_\theta,f_\theta)$ (or $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$) and $\Theta\subseteq\Theta(r)$ (or $\Theta\subseteq\widetilde{\Theta}(r)$) with $r\geq 2$. Then, when the assumption $D(\Theta)$ holds,*

$$\left\|\frac{1}{n}L_n(\theta)-L(\theta)\right\|_\Theta\xrightarrow[n\to+\infty]{a.s.}0\quad\textit{with}\quad L(\theta):=-\frac{1}{2}\mathbb{E}[q_0(\theta)]. \tag{7.20}$$

Proof. See the proof of Theorem 1 in [7]. ■

Lemma 4. *Under the assumptions of Theorem 3.1 and for $m\in\mathcal{M}$ such as $m^*\subset m$, then:*

$$L_n(\widehat{\theta}(m))-L_n(\theta^*)=O_P(1). \tag{7.21}$$

Proof. Applying a second order Taylor expansion of L_n around $\widehat{\theta}(m^*)$ for n sufficiently large such that $\overline{\theta}(m)\in\Theta(m)$ which are between θ^* and $\widehat{\theta}(m^*)$, yields:

$$\begin{aligned}L_n(\widehat{\theta}(m))-L_n(\theta^*)\ &=\ (\widehat{\theta}(m)-\theta^*)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}+\frac{1}{2}(\widehat{\theta}(m)-\theta^*)'\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta\partial\theta'}(\widehat{\theta}(m)-\theta^*)\\&=\sqrt{n}(\widehat{\theta}(m)-\theta^*)\frac{1}{\sqrt{n}}\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta}+\frac{1}{2}\sqrt{n}(\widehat{\theta}(m)-\theta^*)'\frac{1}{n}\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta\partial\theta'}\sqrt{n}(\widehat{\theta}(m)-\theta^*)\\&=\qquad o_p(1)\qquad+\qquad O_P(1)\\&=O_P(1),\end{aligned}$$

by using equality (7.19). ■

7.2. Proof of Theorem 3.1

As we point out in Subsection 2.4, the proof is divided into two parts.

Proof. 1. For $m \in \mathcal{M}$ such as $m* \subset m$ and $m \neq m^*$ (overfitting), then using with $\widehat{C}(m) = -2\widehat{L}_n\big(\widehat{\theta}(m)\big) + |m|\,\kappa_n$ (see (2.8)), we have:

$$\begin{array}{lll} \mathbb{P}(\widehat{m}=m) & \leq & \mathbb{P}\big(\widehat{C}(m) \leq -2\widehat{L}_n\big(\theta^*\big) + |m^*|\,\kappa_n\big) \\ & \leq & \mathbb{P}\Big(-2\big(\widehat{L}_n(\widehat{\theta}) - \widehat{L}_n(\theta^*)\big) \leq \kappa_n(|m^*|-|m|)\Big) \\ & \leq & \mathbb{P}\Big(\dfrac{1}{\kappa_n}\big(\widehat{L}_n(\theta^*) - \widehat{L}_n(\widehat{\theta})\big) \leq \dfrac{(|m^*|-|m|)}{2}\Big) \\ & \underset{n\to\infty}{\longrightarrow} & 0 \end{array}$$

by virtue of Lemma 3 and because $|m|-|m^*| \geq 1$.

2. Let $m \in \mathcal{M}$ such as $m^* \not\subset m$. Then,

$$\widehat{L}_n(\widehat{\theta}(m^*)) - \widehat{L}_n(\widehat{\theta}(m)) = \big(\widehat{L}_n(\widehat{\theta}(m^*)) - L_n(\widehat{\theta}(m^*))\big) - \big(\widehat{L}_n(\widehat{\theta}(m)) - L_n(\widehat{\theta}(m))\big) + \big(L_n(\widehat{\theta}(m^*)) - L_n(\widehat{\theta}(m))\big). \quad (7.22)$$

It follows from Lemma 1 that the first and the second term of the right part of (7.22) are equal to $o_P(\kappa_n)$. Moreover, the third term can be written as follows:

$$L_n(\widehat{\theta}(m^*)) - L_n(\widehat{\theta}(m)) = \big(L_n(\widehat{\theta}(m^*)) - L_n(\theta^*)\big) + \big(L_n(\theta^*) - L_n(\widehat{\theta}(m))\big).$$

From Lemma 4, one deduces $L_n(\widehat{\theta}(m^*)) - L_n(\theta^*) = O_P(1)$. In addition, in the sequel, we are going to show that

$$L_n(\theta^*) - L_n(\widehat{\theta}(m)) = n\,\big(A(m) + o_P(1)\big), \quad \text{with } A(m) > 0. \quad (7.23)$$

For any $\theta \in \Theta(m)$, we have from Proposition 2

$$\begin{aligned} L_n(\theta^*) - L_n(\theta) &= \big(L_n(\theta^*) - n\,L(\theta^*)\big) - \big(L_n(\theta) - n\,L(\theta)\big) + n\,\big(L(\theta^*)) - L(\theta)\big) \\ &= o_P(n) + n\,\big(L(\theta^*)) - L(\theta)\big). \end{aligned}$$

Let us denote by $\mathcal{F}_t := \sigma\big(X_{t-1}, X_{t-2}, \cdots\big)$. Using conditional expectation, we obtain

$$L(\theta^*) - L(\theta) = -\frac{1}{2}\,\mathbb{E}\Big[\mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big]\Big]. \quad (7.24)$$

But,

$$\begin{array}{lll} \mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big] & = & \mathbb{E}\Big[\dfrac{(X_0 - f_\theta^0)^2}{H_\theta^0} + \log(H_\theta^0) - \dfrac{(X_0 - f_{\theta^*}^0)^2}{H_{\theta^*}^0} - \log(H_{\theta^*}^0) \mid \mathcal{F}_0\Big] \\ & = & \log\Big(\dfrac{H_\theta^0}{H_{\theta^*}^0}\Big) + \dfrac{\mathbb{E}\big[(X_0 - f_\theta^0)^2 \mid \mathcal{F}_0\big]}{H_\theta^0} - \dfrac{\mathbb{E}\big[(X_0 - f_{\theta^*}^0)^2 \mid \mathcal{F}_0\big]}{H_{\theta^*}^0} \\ & = & \log\Big(\dfrac{H_\theta^0}{H_{\theta^*}^0}\Big) - 1 + \dfrac{\mathbb{E}\big[(X_0 - f_{\theta^*}^0 + f_{\theta^*}^0 - f_\theta^0)^2 \mid \mathcal{F}_0\big]}{H_\theta^0} \\ & = & \dfrac{H_{\theta^*}^0}{H_\theta^0} - \log\Big(\dfrac{H_{\theta^*}^0}{H_\theta^0}\Big) - 1 + \dfrac{(f_{\theta^*}^0 - f_\theta^0)^2}{H_\theta^0} \end{array}$$

As a consequence, from (7.24),

$$\begin{array}{lll} A(m) & := & 2\,\big(L(\theta^*) - L(\theta)\big) \\ & = & \mathbb{E}\Big[\dfrac{H_{\theta^*}^0}{H_\theta^0} - \log\Big(\dfrac{H_{\theta^*}^0}{H_\theta^0}\Big) - 1 + \dfrac{(f_{\theta^*}^0 - f_\theta^0)^2}{H_\theta^0}\Big] \\ & \geq & \mathbb{E}\Big[\dfrac{H_{\theta^*}^0}{H_\theta^0}\Big] - \log\Big(\mathbb{E}\Big[\dfrac{H_{\theta^*}^0}{H_\theta^0}\Big]\Big) - 1 + \mathbb{E}\Big[\dfrac{(f_{\theta^*}^0 - f_\theta^0)^2}{H_\theta^0}\Big] \quad \text{by Jensen Inequality.} \end{array}$$

Since $x-\log(x)-1>0$ for any $x>0,\ x\neq 1$ and $x-\log(x)-1=0$ for $x=1$, we deduce that

- If $f^0_{\theta^*}\neq f^0_\theta$ then $\mathbb{E}\Big[\frac{(f^0_{\theta^*}-f^0_\theta)^2}{H^0_\theta}\Big]>0$ and $A(m)>0$.
- Otherwise, if $f^0_{\theta^*}=f^0_\theta$, then

  $$A(m)=\mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_\theta}-\log\Big(\frac{H^0_{\theta^*}}{H^0_\theta}\Big)-1\Big],$$

  From Assumption ID(Θ), when $\theta^*\notin\Theta(m)$ and if $f^0_{\theta^*}=f^0_\theta$, we necessarily have $H^0_{\theta^*}\neq H^0_\theta$ so that $\frac{H^0_{\theta^*}}{H^0_\theta}\neq 1$. Then $A(m)>0$.

Therefore $A(m)>0$ for any $\theta\in\Theta(m)$ and particularly for $\theta=\widehat{\theta}(m)$ and (7.23) holds. Thus, (7.22) yields to

$$\widehat{L}_n(\widehat{\theta}(m^*))-\widehat{L}_n(\widehat{\theta}(m))=o_P(\kappa_n)+O_P(1)+n\,A(m)+o_P(n)=O_P(1)+n\,A(m)+o_P(n).$$

Finally, when $m\in\mathcal{M}$ such as $m^*\not\subset m$, we have

$$\widehat{C}(m)-\widehat{C}(m^*)=2\,n\,A(m)+o_P(n)+O_P(1)+\kappa_n(|m|-|m^*|)\ \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}}\ +\infty$$

since $\kappa_n=o(n)$, therefore $\mathbb{P}\big(\widehat{C}(m)>\widehat{C}(m^*)\big)\ \underset{n\to\infty}{\longrightarrow}\ 1$.

Thus we have proved the first and most difficult part of Theorem (3.1). The next lines show the second part which is about the consistency of $\widehat{\theta}(\widehat{m})$.

Given $\epsilon>0$, we have :

$$\begin{aligned}\mathbb{P}\Big(\|\widehat{\theta}(\widehat{m})-\theta^*\|_{i\in m^*}>\epsilon\Big)\quad&=\quad\mathbb{P}\Big(\|\widehat{\theta}(\widehat{m})-\theta^*\|_{i\in m^*}>\epsilon|\widehat{m}=m^*\Big)\,\mathbb{P}\big(\widehat{m}=m^*\big)\\&\qquad+\mathbb{P}\Big(\|\widehat{\theta}(\widehat{m})-\theta^*\|_{i\in m^*}>\epsilon|\widehat{m}\neq m^*\Big)\,\mathbb{P}\big(\widehat{m}\neq m^*\big).\end{aligned}$$

From the strong consistency of the QMLE (see New version of Theorem 1 of [7]), the first term of the right hand side of the above equation is asymptotically zero and also the second one under the assumptions of the first part of Theorem 3.1 which gives $\mathbb{P}\big(\widehat{m}\neq m^*\big)\ \underset{n\to\infty}{\longrightarrow}\ 0$.

■

7.3. Proof of Theorem 3.2

Proof. For $x=(x_i)_{1\leq i\leq d}\in\mathbb{R}^d$, denote $F_n(x)=\mathbb{P}\Big(\bigcap\limits_{1\leq i\leq d}\sqrt{n}\,\big(\widehat{\theta}(\widehat{m})-\theta^*\big)_i\leq x_i\Big)$.

First, we have:

$$\begin{aligned}F_n(x)\quad&=\quad\mathbb{P}\Big(\bigcap_{1\leq i\leq d}\sqrt{n}\,\big(\widehat{\theta}(\widehat{m})-\theta^*\big)_i\leq x_i\ \Big|\ \widehat{m}=m^*\Big)\,\mathbb{P}\big(\widehat{m}=m^*\big)\\&\qquad+\mathbb{P}\Big(\bigcap_{1\leq i\leq d}\sqrt{n}\,\big(\widehat{\theta}(\widehat{m})-\theta^*\big)_i\leq x_i\ \Big|\ \widehat{m}\neq m^*\Big)\,\mathbb{P}\big(\widehat{m}\neq m^*\big).\end{aligned}$$

Under the assumptions of Theorem 3.1, $\mathbb{P}\big(\widehat{m}=m^*\big)\ \underset{n\to\infty}{\longrightarrow}\ 1$ and $\mathbb{P}\big(\widehat{m}\neq m^*\big)\ \underset{n\to\infty}{\longrightarrow}\ 0$. Therefore the second term in the right side of the previous equality asymptotically vanishes. For the first term, we can write,

$$\begin{aligned}&\mathbb{P}\Big(\bigcap_{1\leq i\leq d}\sqrt{n}\,\big(\widehat{\theta}(\widehat{m})-\theta^*\big)_i\leq x_i\ \Big|\ \widehat{m}=m^*\Big)\\&\qquad=\mathbb{P}\Big(\Big\{\bigcap_{i\in m^*}\sqrt{n}\,\big(\widehat{\theta}(m^*)-\theta^*\big)_i\leq x_i\Big\}\bigcap\Big\{\bigcap_{i\notin m^*}\sqrt{n}\,\big(\widehat{\theta}(m^*)-\theta^*\big)_i\leq x_i\Big\}\Big).\end{aligned}$$

Since $\theta(m^*) \in \Theta(m^*)$, $\big((\widehat{\theta}(m^*))_i\big)_{i \notin m^*} = \big(\theta^*_i\big)_{i\notin m^*} = 0$, for $(x_i)_{i\notin m^*}$ a family of non negative real numbers we have:

$$\begin{aligned}\mathbb{P}\Big(\Big\{ \bigcap_{i\in m^*} \sqrt{n}\,\big(\widehat{\theta}(m^*) - \theta^*\big)_i \leq x_i\Big\} \bigcap \Big\{ \bigcap_{i\notin m^*} \sqrt{n}\,\big(\widehat{\theta}(m^*) - \theta^*\big)_i \leq x_i\Big\}\Big) \\ = \mathbb{P}\Big(\bigcap_{i\in m^*} \sqrt{n}\,\big(\widehat{\theta}(m^*) - \theta^*\big)_i \leq x_i\Big) \\ \underset{n\to\infty}{\longrightarrow} \mathbb{P}\Big(\big(F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}\big)^{-1/2} Z \leq (x_i)_{i\in m^*}\Big),\end{aligned}$$

with Z a standard Gaussian random vector in $\mathbb{R}^{|m^*|}$ from the central limit theorem (3.3), and this achieves the proof of 3.5 of Theorem 3.2. ■

7.4. Proof of Theorem 5.1

Consider the following notation: for $\theta \in \Theta$ and $m \in \mathcal{M}$, denote the residuals and quasi-residuals by:

$$\begin{cases} e_t(\theta) \;:=\; \big(M^t_\theta\big)^{-1}\big(X_t - f^t_\theta\big) \quad \text{and} \quad \widehat{e}_t(\theta) \;:=\; \big(\widehat{M}^t_\theta\big)^{-1}\big(X_t - \widehat{f}^t_\theta\big) \\ e_t(m) \;:=\; \big(M^t_{\widehat{\theta}(m)}\big)^{-1}\big(X_t - f^t_{\widehat{\theta}(m)}\big) \quad \text{and} \quad \widehat{e}_t(m) \;:=\; \big(M^t_{\widehat{\theta}(m)}\big)^{-1}\big(X_t - \widehat{f}^t_{\widehat{\theta}(m)}\big)\end{cases}.$$

For $k \in \{0,1,\dots,n-1\}$, $\theta\in\Theta$ and $m \in \mathcal{M}$, define also the adjusted lag-k covariograms and correlograms of the squared (standardized) residual by:

$$\begin{cases} \gamma_k(\theta) := \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k} \big(e^2_t(\theta) - 1\big)\big(e^2_{t+k}(\theta)-1\big) \quad \text{and} \quad \widehat{\gamma}_k(\theta) := \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k} \big(\widehat{e}^2_t(\theta) - 1\big)\big(\widehat{e}^2_{t+k}(\theta)-1\big) \\ \gamma_k(m) := \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k} \big(e^2_t(m) - 1\big)\big(e^2_{t+k}(m)-1\big) \quad \text{and} \quad \widehat{\gamma}_k(m) := \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k} \big(\widehat{e}^2_t(m) - 1\big)\big(\widehat{e}^2_{t+k}(m)-1\big)\end{cases}$$

and $\rho_k(\theta) := \dfrac{\gamma_k(\theta)}{\gamma_0(\theta)}$, $\widehat{\rho}_k(\theta) := \dfrac{\widehat{\gamma}_k(\theta)}{\widehat{\gamma}_0(\theta)}$, $\rho_k(m) := \dfrac{\gamma_k(m)}{\gamma_0(m)}$ and $\widehat{\rho}_k(m) := \dfrac{\widehat{\gamma}_k(m)}{\widehat{\gamma}_0(m)}$.

Finally, for K a positive integer, denote the vector of adjusted correlogram:

$$\widehat{\rho}(\theta) := \big(\widehat{\rho}_1(\theta),\dots,\widehat{\rho}_K(\theta)\big)' \quad \text{and} \quad \widehat{\rho}(m) := \big(\widehat{\rho}_1(m),\dots,\widehat{\rho}_K(m)\big)'.$$

Proof. (1) This proof is divided into two parts. In (i) we prove a result that ensures that the asymptotic distributions of the vectors $\widehat{\rho}(\theta)$ and $\rho(\theta)$ are the same. In (ii) we show that the large sample distribution of $\sqrt{n}\rho(m^*)$ is normal with a covariance matrix $V(\theta^*, m^*)$. Those two conditions do lead well to the asymptotic normality (5.1).

(i) In this part, we first show that for any $k \in \mathbb{N}$,

$$\sqrt{n}\,\big\|\widehat{\gamma}_k(\theta) - \gamma_k(\theta)\big\|_\Theta \xrightarrow[n\to\infty]{a.s.} 0. \tag{7.25}$$

We have:

$$\begin{aligned}\sqrt{n}\big(\widehat{\gamma}_k(\theta) - \gamma_k(\theta)\big) &= \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n}\big(\widehat{e}^2_t(\theta) - 1\big)\big(\widehat{e}^2_{t-k}(\theta)-1\big) - \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n}\big(e^2_t(\theta) - 1\big)\big(e^2_{t-k}(\theta)-1\big) \\ &= \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n}\big(\widehat{e}^2_t(\theta)\widehat{e}^2_{t-k}(\theta) - e^2_t(\theta)e^2_{t-k}(\theta)\big) + \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n}\big(\widehat{e}^2_t(\theta) - e^2_t(\theta)\big) \\ &\qquad + \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n}\big(e^2_{t-k}(\theta) - \widehat{e}^2_{t-k}(\theta)\big) \\ &=: I_1 + I_2 + I_3.\end{aligned}$$

Now, we show that $\|I_1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$. We can rewrite I_1 as follows

$$\begin{aligned} I_1 &= \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n} \widehat{e}^2_{t-k}(\theta)\big(\widehat{e}^2_t(\theta)-e^2_t(\theta)\big) + \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n} e^2_t(\theta)\big(\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big) \\ &= \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n} \big(\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big)\big(\widehat{e}^2_t(\theta)-e^2_t(\theta)\big) + \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n} e^2_{t-k}(\theta)\big(\widehat{e}^2_t(\theta)-e^2_t(\theta)\big) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \frac{1}{\sqrt{n}}\sum_{t=k+1}^{n} e^2_t(\theta)\big(\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big) \\ &:= I_1^1 + I_1^2 + I_1^3. \end{aligned}$$

Let us show that $\|I_1^1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ in our two frameworks.
a/ If $X \subset AC(M_\theta, f_\theta)$, by Hölder's inequality, it follows from (7.8) that,

$$\mathbb{E}\Big[\Big\|\big(\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big)\big(\widehat{e}^2_t(\theta)-e^2_t(\theta)\big)\Big\|_\Theta^{1/2}\Big] \quad\le\quad \Big(\mathbb{E}\big[\big\|\widehat{e}^2_t(\theta)-e^2_t(\theta)\big\|_\Theta\big]\times\mathbb{E}\big[\big\|\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big\|_\Theta\big]\Big)^{1/2}.$$

But we have

$$\big\|\widehat{e}^2_t(\theta)-e^2_t(\theta)\big\|_\Theta \le \frac{1}{\underline{h}}\,\big(2|X_t| + \|\widehat{f}^t_\theta\|_\Theta + \|f^t_\theta\|_\Theta\big)\big\|\widehat{f}^t_\theta - f^t_\theta\|_\Theta + \frac{4}{\underline{h}^{3/2}}\,\big(|X_t|^2 + \|f^t_\theta\|^2_\Theta\big)\big\|\widehat{M}^t_\theta - M^t_\theta\|_\Theta.$$

Therefore,

$$\begin{aligned} \mathbb{E}\big[\big\|\widehat{e}^2_t(\theta)-e^2_t(\theta)\big\|_\Theta\big] \quad&\le\quad C\,\Big(\mathbb{E}\big[\big(|X_t|^2 + \|\widehat{f}^t_\theta\|^2_\Theta + \|f^t_\theta\|^2_\Theta\big)\big]\;\times\;\mathbb{E}\big[\|\widehat{f}^t_\theta - f^t_\theta\|^2_\Theta\big]\Big)^{1/2} \\ &\qquad + C\,\Big(\mathbb{E}\big[\big(|X_t|^4 + \|f^t_\theta\|^2_\Theta\big)\big]\;\times\;\mathbb{E}\big[\|\widehat{M}^t_\theta - M^t_\theta\|^2_\Theta\big]\Big)^{1/2} \\ &\le\quad C\,\Big(\mathbb{E}\Big[\Big|\sum_{j\ge t}\alpha_j(f_\theta,\Theta)X_{t-j}\Big|^2\Big]\Big)^{1/2} + C\,\Big(\mathbb{E}\Big[\Big|\sum_{j\ge t}\alpha_j(M_\theta,\Theta)X_{t-j}\Big|^2\Big]\Big)^{1/2} \\ &\le\quad C\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta), \end{aligned}$$

using $\mathbb{E}\big[|X_t|^4 + \|f^t_\theta\|^2_\Theta + \|\widehat{f}^t_\theta\|^2_\Theta\big] < \infty$ and Cauchy-Schwarz Inequality. Hence,

$$\mathbb{E}\Big[\Big\|\big(\widehat{e}^2_{t-k}(\theta)-e^2_{t-k}(\theta)\big)\big(\widehat{e}^2_t(\theta)-e^2_t(\theta)\big)\Big\|_\Theta^{1/2}\Big] \quad\le\quad C\sum_{j\ge t-k}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta).$$

Therefore, from [27], $\|I_1^1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ when

$$\sum_{t=1}^{\infty} t^{-1/4}\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta) < \infty. \tag{7.26}$$

b/ if $X \subset \widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$, same computations imply $\|I_1^1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ when

$$\sum_{t=1}^{\infty} t^{-1/4}\sum_{j\ge t}\alpha_j(\widetilde{H}_\theta,\Theta) < \infty. \tag{7.27}$$

Since $\mathbb{E}\big[\|e^2_t(\theta)\|_\Theta\big] \le 2\,\underline{h}^{-1}\mathbb{E}\big[X_t^2 + \|f^t_\theta\|^2_\Theta\big] < \infty$ and similarly $\mathbb{E}\big[\|\widehat{e}^2_t(\theta)\|_\Theta\big] < \infty$, we deduce from the same inequalities as in the first case of I_1^1 that $\|I_1^2\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ and $\|I_1^3\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ when

$$\sum_{t=1}^{\infty} t^{-1/4}\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\widetilde{H}_\theta,\Theta)\Big)^{1/2} < \infty, \tag{7.28}$$

which is also the condition for insuring that $\|I_2\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ and $\|I_3\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$. This ends the proof of (7.25).
Finally, since $\widehat{\rho}_k(\theta)=\widehat{\gamma}_k(\theta)/\widehat{\gamma}_0(\theta)$ and $\rho_k(\theta)=\gamma_k(\theta)/\gamma_0(\theta)$, with $\gamma_0(\theta)>0$, we deduce under condition (7.28) that

$$\sqrt{n}\big\|\widehat{\rho}_k(\theta)-\rho_k(\theta)\big\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0 \quad \text{for any } k\geq 1. \tag{7.29}$$

This also implies

$$\sqrt{n}\big|\widehat{\rho}_k(m^*)-\rho_k(m^*)\big| \xrightarrow[n\to+\infty]{a.s.} 0 \quad \text{for any } k\geq 1. \tag{7.30}$$

(ii) The proof of this result has already been done in [31] but in a Gaussian framework. We recall here the main lines while avoiding the Gaussian assumption. The first step is to use a Taylor expansion of the function γ. Hence, we have for each $k=1,\dots,K$,

$$\sqrt{n}\,\gamma_k(m^*)=\sqrt{n}\,\gamma_k(\widehat{\theta}(m^*))=\sqrt{n}\,\gamma_k(\theta^*)+\partial_\theta\gamma_k(\overline{\theta}^{(k)})\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}, \tag{7.31}$$

where $\partial_\theta\gamma_k={}^t\big(\partial\gamma_k/\partial\theta_i\big)_{i\in m^*}$, and $\overline{\theta}^{(k)}$ is in the ball of centre θ^* and radius $\|(\widehat{\theta}(m^*)-\theta^*)_{i\in m^*}\|$. We also have

$$\begin{aligned}\partial_\theta\gamma_k(\theta)=-\frac{2}{n}\Big(\sum_{t=k+1}^{n} e_t^2(\theta)\,\big(e_{t-k}^2(\theta)-1\big)\frac{\partial_\theta M_\theta^t}{M_\theta^t}+e_t(\theta)\big(e_{t-k}^2(\theta)-1\big)\,\frac{\partial_\theta f_\theta^t}{M_\theta^t}\\ +\,e_{t-k}(\theta)\,\big(e_t^2(\theta)-1\big)\frac{\partial_\theta f_\theta^{t-k}}{M_\theta^{t-k}}+e_{t-k}^2(\theta)\,\big(e_t^2(\theta)-1\big)\frac{\partial_\theta M_\theta^{t-k}}{M_\theta^{t-k}}\Big).\end{aligned} \tag{7.32}$$

We have $\mathbb{E}\big[e_{t-k}(\theta^*)\,\big(e_t^2(\theta^*)-1\big)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}}\mid\sigma\big((\xi_s)_{s\leq t-k}\big)\big]=e_{t-k}(\theta^*)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}}\mathbb{E}\big[e_t^2(\theta^*)-1\big]=0$ since we have assumed $\mathbb{E}[\xi_0^2]=1$. Moreover, $\mathbb{E}\big[e_t(\theta^*)\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big]=\mathbb{E}\big[\xi_t\,\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big]=0$ and this implies $\mathbb{E}\big[e_t(\theta^*)\big(e_{t-k}^2(\theta^*)-1\big)\,\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big]=0$. As a consequence, the expectation of the three last terms of (7.32) vanishes for $\theta=\theta^*$. By using the Ergodic Theorem, we finally obtained:

$$\partial_\theta\gamma_k(\theta^*) \xrightarrow[n\to+\infty]{a.s.} -2\,\mathbb{E}\Big[e_k^2(\theta^*)\,\big(e_0^2(\theta^*)-1\big)\frac{\partial_\theta M_{\theta^*}^k}{M_{\theta^*}^k}\Big]=-2\,\mathbb{E}\Big[\big(\xi_0^2-1\big)\,\partial_\theta\log\big(M_{\theta^*}^k\big)\Big].$$

Moreover, since $\partial^2_{\theta^2}f_\theta$ and $\partial^2_{\theta^2}M_\theta$ exist, and since $\widehat{\theta}(m^*) \xrightarrow[n\to+\infty]{a.s.} \theta^*$, we deduce that the same almost sure convergence occurs for $\partial_\theta\gamma_k(\overline{\theta}^{(k)})$. Then, we finally obtain

$$\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_{1\leq k\leq K} \xrightarrow[n\to+\infty]{a.s.} J_K(m^*)=-2\,\Big(\mathbb{E}\Big[\big(\xi_0^2-1\big)\,\frac{\partial}{\partial\theta_j}\log\big(M_{\theta^*}^i\big)\Big]\Big)_{1\leq i\leq K,\,j\in m^*}. \tag{7.33}$$

We also established a central limit theorem for $\widehat{\theta}(m^*)$ in (3.3), and this implies

$$\begin{aligned}&\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_{1\leq k\leq K}\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}\\ &\qquad\xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\Big(0\;,\;J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J'_K(m^*)\Big).\end{aligned} \tag{7.34}$$

On the other hand, when $\theta=\theta^*$, $e_t^2(\theta^*)=\xi_t^2$ for any $t\in\mathbb{Z}$ and since $\mathbb{E}[\xi_0^2]=1$, we deduce that $\big(e_t^2(\theta^*)-1\big)_t$ is a sequence of centred iid random variables with variance μ_4-1 with $\mu_4=\mathbb{E}[\xi_0^4]$. In such as case, the asymptotic behavior of the covariograms is well known and we deduce:

$$\sqrt{n}\,\big(\gamma_k(\theta^*)\big)_{1\leq k\leq K} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\big(0\,,\,(\mu_4-1)^2\,I_K\big), \tag{7.35}$$

with I_k the $(K\times K)$ identity matrix.

We would like to use (7.31) for obtaining the asymptotic behavior of $\gamma(m^*)$. In (7.34) and (7.35), we obtained the asymptotic normality of each of the two terms composing $\gamma(m^*)$. Now we need to study the joint asymptotic behavior of $\sqrt{n}\,\gamma(\theta^*)$ and $\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}$.

Using the proof of the asymptotic normality of the QMLE (see for instance [7]), a Taylor expansion of log-likelihood for large n leads to

$$\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}\approx -\big(F(\theta^*,m^*)\big)^{-1}\frac{1}{n}\,\frac{\partial}{\partial\theta}L_n(\theta^*).$$

Therefore, the asymptotic cross expectation between $\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_k\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}$ and $\sqrt{n}\,\gamma(\theta^*)$ is equal to:

$$-\;J_K(m^*)\,F(\theta^*,m^*)^{-1}\mathbb{E}\Big[\frac{\partial}{\partial\theta}L_n(\theta^*)\,\gamma(\theta^*)'\Big]. \tag{7.36}$$

From (2.5), a direct differentiation of L_n provides

$$\frac{\partial}{\partial\theta}L_n(\theta^*)=\sum_{t=1}^n\big(e_t^2(\theta^*)-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^t_{\theta^*}\big)+\sum_{t=1}^n e_t(\theta^*)\,\frac{\partial}{\partial\theta}f^t_{\theta^*}$$

so that,

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial}{\partial\theta}L_n(\theta^*)\,\gamma_k(\theta^*)\Big]\;&=\;\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^n\big(e_i^2(\theta^*)-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^i_{\theta^*}\big)\sum_{j=k+1}^n\big(e_j^2(\theta^*)-1\big)\,\big(e_{j-k}^2(\theta^*)-1\big)\Big]\\&\qquad+\frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^n e_i(\theta^*)\,\frac{\partial}{\partial\theta}f^i_{\theta^*}\sum_{j=k+1}^n\big(e_j^2(\theta^*)-1\big)\,\big(e_{j-k}^2(\theta^*)-1\big)\Big]\\&=\;\frac{1}{n}\sum_{i=1}^n\sum_{j=k+1}^n\mathbb{E}\Big[\big(\xi_i^2-1\big)\,\big(\xi_j^2-1\big)\,\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^i_{\theta^*}\big)\Big]\\&\qquad+\frac{1}{n}\sum_{i=1}^n\sum_{j=k+1}^n\mathbb{E}\Big[\xi_i\,\big(\xi_j^2-1\big)\,\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}f^i_{\theta^*}\Big].\end{aligned}$$

Using conditional expectations, we have $\mathbb{E}\Big[\big(\xi_i^2-1\big)\,\big(\xi_j^2-1\big)\,\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^i_{\theta^*}\big)\Big]=0$ for $i\neq j$ since $k\geq 1$. Moreover, for $i=j$, we obtain:

$$\mathbb{E}\Big[\big(\xi_i^2-1\big)\,\big(\xi_j^2-1\big)\,\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^i_{\theta^*}\big)\Big]=(\mu_4-1)\,\mathbb{E}\Big[\big(\xi_{i-k}^2-1\big)\,\frac{\partial}{\partial\theta}\log\big(M^i_{\theta^*}\big)\Big],$$

which is the row k of matrix $-\frac{(\mu_4-1)}{2}\,J_K(m^*)$. Similarly, and using the assumption $\mathbb{E}\big[\xi_0^3\big]=0$, we obtain $\mathbb{E}\Big[\xi_i\,\big(\xi_j^2-1\big)\,\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}f^i_{\theta^*}\Big]=0$ for any i,j and k. As a consequence,

$$\begin{aligned}&\mathrm{Cov}\,\Big(\sqrt{n}\,\gamma(\theta^*)\,,\,\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_k\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta^*_i\big)_{i\in m^*}\Big)\\&\qquad\qquad\underset{n\to\infty}{\longrightarrow}\;\frac{1}{2}\,(\mu_4-1)\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\,J'_K(m^*).\end{aligned}$$

Finally, we deduce the asymptotic covariance matrix of $\sqrt{n}\,\gamma(m^*)$, which is

$$\begin{aligned}&(\mu_4-1)^2\,I_K+J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J'_K(m^*)\\&\qquad\qquad+(\mu_4-1)\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\,J'_K(m^*).\end{aligned}$$

Moreover the vector $\gamma(m^*)$ is normal distributed from Lemma 3.3 of [32].
Thus, using Slutsky Lemma and with $\gamma_0(m^*)\;\underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}}\;\mu_4-1$, and with $\rho_k(m^*)=\gamma_k(m^*)/\gamma_0(m^*)$, the limit theorem (5.1) holds with

$$\begin{aligned}V(\theta^*,m^*):=I_K+(\mu_4-1)^{-2}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J'_K(m^*)&\\+(\mu_4-1)^{-1}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\,J'_K(m^*).&\end{aligned}\tag{7.37}$$

The proof is achieved after using the limit theorem (7.30).

(2) (5.2) follows directly from (5.1).

(3) We follow a same reasoning like in the proof of Theorem 3.2. For $x = (x_k)_{1 \leq k \leq K} \in \mathbb{R}^K$, denote by $F_n(x) = \mathbb{P}\Big(\bigcap_{1 \leq k \leq K} \sqrt{n} \left(\widehat{\rho}(\widehat{m}) \right)_k \leq x_k \Big)$ the distribution function of $\sqrt{n} \widehat{\rho}(\widehat{m})$.
Applying the Total Probability Rule and by virtue of Theorem 3.1, we obtain:

$$F_n(x) = \mathbb{P}\Big(\bigcap_{1 \leq k \leq K} \sqrt{n} \left(\widehat{\rho}(m^*) \right)_k \leq x_k \Big).$$

Therefore, the vectors $\sqrt{n} \widehat{\rho}(\widehat{m})$ and $\sqrt{n} \widehat{\rho}(m^*)$ have exactly the same distribution. ■

References

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd international symposium on information, Akademiai Kiado, Budapest* (1973).

[2] Allen, D. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics 16*, 1 (1974), 125–127.

[3] Alquier, P., and Wintenberger, O. Model selection for weakly dependent time series forecasting. *Bernoulli 18*, 3 (2012), 883–913.

[4] Arkoun, O., Brua, J.-Y., and Pergamenshchikov, S. Sequential model selection method for nonparametric autoregression. *arXiv preprint arXiv:1809.02241* (2018).

[5] Arlot, S., and Massart, P. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research 10* (2009), 245–279.

[6] Bardet, J.-M., Boularouk, Y., and Djaballah, K. Asymptotic behavior of the laplacian quasi-maximum likelihood estimator of affine causal processes. *Electronic journal of statistics 11*, 1 (2017), 452–479.

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

[8] Berkes, I., Horváth, L., and Kokoszka, P. GARCH processes: structure and estimation. *Bernoulli 9* (2003), 201–227.

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

[10] Ding, J., Tarokh, V., and Yang, Y. Bridging aic and bic: a new criterion for autoregression. *IEEE Transactions on Information Theory 64*, 6 (2018), 4024–4043.

[11] Ding, J., Tarokh, V., and Yang, Y. Model selection techniques: An overview. *IEEE Signal Processing Magazine 35*, 6 (2018), 16–34.

[12] Ding, Z., Granger, C., and Engle, R. A long memory property of stock market returns and a new model. *Journal of empirical finance 1*, 1 (1993), 83–106.

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

[14] Duchesne, P., and Francq, C. On diagnostic checking time series models with portmanteau test statistics based on generalized inverses and. In *COMPSTAT 2008*. Springer, 2008, pp. 143–154.

[15] Francq, C., and Zakoïan, J.-M. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli 10* (2004), 605–637.

[16] Gao, J., and Tong, H. Semiparametric non-linear time series model selection. *Journal of the Royal Statistical Society: Series B 66*, 2 (2004), 321–336.

[17] Hannan, E. The estimation of the order of an arma process. *The Annals of Statistics 8*, 5 (1980), 1071–1081.

[18] Hoerl, A., and Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[19] Hsu, H.-L., Ing, C.-K., and Tong, H. On model selection from a finite family of possibly misspecified time series models. *The Annals of Statistics 47*, 2 (2019), 1061–1087.

[20] Hurvich, C., and Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika 76*, 2 (1989), 297–307.

[21] Ing, C.-K. Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics 35*, 3 (2007), 1238–1277.

[22] Ing, C.-K., Sin, C.-Y., and Yu, S.-H. Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis 106* (2012), 57–71.

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

[24] Jeantheau, T. Strong consistency of estimators for multivariate arch models. *Econometric Theory 14*, 1 (1998), 70–86.

[25] Kapetanios, G. Model selection in threshold models. *Journal of Time Series Analysis 22*, 6 (2001), 733–754.

[26] Kock, A. Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory 32*, 1 (2016), 243–259.

[27] Kounias, E., and Weng, T. An inequality and almost sure convergence. *The Annals of Mathematical Statistics 40*, 3 (1969), 1091–1093.

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

[29] Li, G., and Li, W. Least absolute deviation estimation for fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity. *Biometrika 95*, 2 (2008), 399–414.

[30] Li, W. On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika 79*, 2 (1992), 435–437.

[31] Li, W., and Mak, T. On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Time Series Analysis 15*, 6 (1994), 627–636.

[32] Ling, S., and Li, W.-K. Diagnostic checking of nonlinear multivariate time series with multivariate arch errors. *Journal of Time Series Analysis 18*, 5 (1997), 447–464.

[33] Ling, S., and McAleer, M. Asymptotic theory for a vector arma-garch model. *Econometric theory 19*, 2 (2003), 280–310.

[34] Mallows, C. Some comments on cp. *Technometrics 15*, 4 (1973), 661–675.

[35] McQuarrie, A., and Tsai, C. *Regression and Time Series Model Selection.* World Scientific Pub Co Inc, 1998.

[36] Rao, C., Wu, Y., Konishi, S., and Mukerjee, R. On model selection. *Lecture Notes-Monograph Series* (2001), 1–64.

[37] Ren, Y., and Zhang, X. Subset selection for vector autoregressive processes via adaptive lasso. *Statistics & probability letters 80*, 23-24 (2010), 1705–1712.

[38] Schwarz, G. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.

[39] Shao, Q., and Yang, L. Oracally efficient estimation and consistent model selection for autoregressive moving average time series with trend. *Journal of the Royal Statistical Society: Series B 79*, 2 (2017), 507–524.

[40] Shi, P., and Tsai, C.-L. Regression model selection-a residual likelihood approach. *Journal of the Royal Statistical Society: Series B 64*, 2 (2002), 237–252.

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

[42] Sin, C.-Y., and White, H. Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics 71*, 1-2 (1996), 207–225.

[43] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B* (1974), 111–147.

[44] Straumann, D., and Mikosch, T. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics 34*, 5 (2006), 2449–2495.

[45] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* (1996), 267–288.

[46] Tsay, R. Order selection in nonstationary autoregressive models. *The Annals of Statistics 12*, 4 (1984), 1425–1433.

[47] Tse, Y., and Zuo, X. Testing for conditional heteroscedasticity: Some monte carlo results. *Journal of Statistical Computation and Simulation 58*, 3 (1997), 237–253.

[48] White, H. Maximum likelihood estimation of misspecified models. *Econometrica* (1982), 1–25.

[49] ZOU, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 476 (2006), 1418–1429.

[50] ZOU, H., AND HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*, 2 (2005), 301–320.