Electronic Journal of Statistics
Vol. 14 (2020) 2009–2052
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1709

Consistent model selection criteria and goodness-of-fit test for common time series models

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; kamilakare@gmail.com

William Kengne[†]

CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France.
e-mail: william.kengne@u-cergy.fr

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. We also propose a tool for diagnosing the goodness-of-fit of the chosen model based on a Portmanteau test. Monte-Carlo experiments and numerical applications on illustrative examples are performed to highlight the obtained asymptotic results. Moreover, using a data-driven choice of the penalty, they show the practical efficiency of this new model selection procedure and Portemanteau test.

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

Received July 2019.

Contents

[*] 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.

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 ([50], [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 best The idea of penalizing dates back to the 1970s with the works of [41] and [1]. Although it is likely that these ideas have already existed in other contexts such as subset selection by [11], [24], and ridge regression by [25].

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, [45] 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 [25], the $\mathbb{L}^1$

norm used by [52] that provides the LASSO procedure and the elastic-net that mixes the $\mathbb{L}^1$ and $\mathbb{L}^2$ norms [57].

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, we are interested to construct a consistent procedure for the general class of times series known as *affine causal processes*, which includes the most common time series.

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 $M\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sqrt{a_0 + a_1 X_{t-1}^2 + \cdots + a_p X_{t-p}^2}$ and $f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = 0$, then $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process.

Numerous classical time series models such as ARMA(p, q), GARCH(p, q), ARMA(p, q)-GARCH(p, q) (see [16] and [40]) or APARCH(δ, p, q) processes (see [16]) belongs to $\mathcal{AC}(M, f)$. The existence of stationary and ergodic solutions of this class has been studied in [17] and [9].

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 [42] 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 consistency assumes that the true model exists and it is included in the collection of candidate models; while the efficiency does not necessarily require the existence of a true model. 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 [23] or [53]; and the asymptotic efficiency of the AIC has been proved, see, among others, [48], [27] for a corrected version of AIC for small samples, [30], [28], [29] for the case of infinite order autoregressive model. [47] propose the (consistent) residual information criteria (RIC) for regression model (including regression models with ARMA errors) selection. In the framework of nonlinear threshold models, [32] proved consistency results of a large class of information criteria, whereas [21] focused 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 [43], 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 [15].

The adaptive lasso, introduced by [56] for variable selection in linear regression models has been extended by [44] to vector autoregressive models, [33] carried out this procedure in stationary and nonstationary autoregressive models; the oracle efficient is established. [35] considers model selection for density estimation under mixing conditions and derived oracle inequalities of the slope heuristic procedure ([14] or [6]); whereas [3] develop oracle inequalities for model selection for weakly dependent time series forecasting. Recently, [46] 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. [26] 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 collection $\mathcal{M}$?
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.

(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, a similar test has been proposed by [38] under the Gaussian assumption on the observations, whereas [39] focused for multivariate time series with multivariate ARCH-type errors. These papers are also based on exact likelihood estimators that do not make feasible Portemanteau tests. [18] proposed an interesting Portmanteau test statistic directly based on the autocorrelations of residuals (and not squared residuals) computed from quasi-likelihood estimators 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.

Monte-Carlo experiments and numerical applications on illustrative examples are also performed to highlight the obtained asymptotic results. We have considered a data-driven choice of the penalty obtained from the slope heuristic procedure (see for instance [6]) for avoiding an a priori choice of the penalty sequence. The simulation study and real data applications show that the results of the proposed model selection procedure and Portetemanteau test are overall satisfactory.

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 studied in Section 5. Finally, numerical results are presented in Section 6 and Section 7 contains the proofs.

2. General Framework

In this section, we are going to present the model selection using Gaussian quasi-maximum likelihood estimators (QMLE) and give some notations in order to facilitate the presentation.

2.1. Quasi-maximum likelihood estimation and model selection

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 QMLE of θ for each specific model m.

This approach as semi-parametric estimation has been successively introduced for GARCH(p, q) processes in [31] where its consistency is also proved, and the asymptotic normality of this estimator has been established in [12] and [19]. In [9], those results have been extended to affine causal processes, and an extension to Laplacian QMLE has been also proposed in [7].

The Gaussian QMLE is derived from the conditional (with respect to the filtration $\sigma\{(X_t)_{t\leq 0}\}$) 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:

$$L_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} q_t(\theta) \ , \ \text{with } q_t(\theta) := \frac{(X_t - f_\theta^t)^2}{H_\theta^t} + \log(H_\theta^t) \tag{2.1}$$

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 unobserved, 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 (2.3) even if the white noise is not Gaussian. Hence, the conditional Gaussian quasi log-likelihood (up to an additional constant) is given for all $\theta \in \Theta$ by

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

$$\text{where} \left\{ \begin{array}{lcl} \widehat{f}_\theta^t & := & f_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{M}_\theta^t & := & M_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{H}_\theta^t & := & (\widehat{M}_\theta^t)^2 \end{array} \right. \tag{2.2}$$

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

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

$$\widehat{\theta}(m) = \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\ \widehat{L}_n(\theta). \tag{2.3}$$

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

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \widehat{C}(m) \quad \text{with} \quad \widehat{C}(m) = -2\widehat{L}_n\big(\widehat{\theta}(m)\big) + |m|\,\kappa_n, \tag{2.4}$$

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.*

$$\mathbb{P}(\widehat{m} = m^*) \underset{n\to\infty}{\longrightarrow} 1; \tag{2.5}$$

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^*.

2.2. The affine causal framework

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 $\mathbb{R}^\infty$.

Before diving in details, let's give some notations that will be useful throughout the paper. 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 $\|g\|_\Theta = \sup\limits_{\theta\in\Theta}\big\{\|g(\theta)\|\big\}$.

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 a model belonging to a collection $\mathcal{M}$, for instance a set of ARMA(p,q) and GARCH(p',q') processes for $0\leq p\leq p_{\max}$, $0\leq q\leq q_{\max}$, $0\leq p'\leq p'_{\max}$, $0\leq q'\leq q'_{\max}$ (where $p_{\max}, q_{\max}, p'_{\max}, q'_{\max}$ are the upper bounds of orders). Then, we would like to chose in this family a "best" model for fitting the data $(X_1,\ldots,X_n)$. For instance, if $p_{\max}=q_{\max}=p'_{\max}=q_{\max}=9$, in the collection above, there is 200 possible models and we expect to recognize the true process (which is unknown to the analyst) 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 $(X_1,\dots,X_n)$ 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:

$$X_t=M_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t+f_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\ \text{ for any } t\in\mathbb{Z}. \tag{2.6}$$

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

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 $\mathcal{M}_1$ such as $\mathcal{M}_1=\big\{\{1\},\{1,2\},\dots,\{1,\dots,q+1\}\big\}$: this family is the hierarchical one of ARCH processes with orders varying from 0 to q.
• a family $\mathcal{M}_2$ such as $\mathcal{M}_2=\mathcal{P}\big(\{1,\dots,p+q+1\}\big)$: this family is the exhaustive one and contains as well the AR(2) process $X_t=\phi_2X_{t-2}+\theta_0\,\xi_t$ as the process $X_t=\phi_1X_{t-1}+\phi_3X_{t-3}+\xi_t\sqrt{\theta_0+a_2X^2_{t-2}}$.

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 [9], 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:

$$X_t = \xi_t \sqrt{\widetilde{H}_\theta\big((X^2_{t-i})_{i\in\mathbb{N}^*}\big)} \ \text{ for any } t\in\mathbb{Z}. \tag{2.7}$$

Therefore, if $M^2_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = H_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \widetilde{H}_\theta\big((X^2_{t-i})_{i\in\mathbb{N}^*}\big)$ 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:

$$\widetilde{\Theta}(r) = \Big\{\theta\in\mathbb{R}^d,\ A(\widetilde{H}_\theta,\{\theta\}) \text{ holds with } \big(\|\xi_0\|_r\big)^2 \sum_{k=1}^{\infty} \alpha_k(\widetilde{H}_\theta,\{\theta\}) < 1\Big\}. \tag{2.8}$$

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 $\|X_0\|_r<\infty$.

3. Asymptotic results

3.1. Assumptions required for the asymptotic study

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 [17], 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:*

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

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

$$\Theta(r) = \Big\{\theta\in\mathbb{R}^d,\ A(f_\theta,\{\theta\}) \text{ and } A(M_\theta,\{\theta\}) \text{ hold with} \\ \sum_{k=1}^{\infty}\alpha_k(f_\theta,\{\theta\}) + \|\xi_0\|_r \sum_{k=1}^{\infty}\alpha_k(M_\theta,\{\theta\}) < 1\Big\}. \tag{3.1}$$

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 [17] and [9]).

Secondly, note that the definitions of the conditional log-likelihood (2.1) and quasi log-likelihood (2.2) 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 $\inf\limits_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ for all $x \in \mathbb{R}^\infty$.*

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

Assumption Id(Θ): *For all θ, $\theta' \in \Theta$, $\big(f^0_\theta = f^0_{\theta'}$ and $M^0_\theta = M^0_{\theta'}\big)$ 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 [9], [51], [55]). 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(Θ): For any $\theta \in \Theta$, $\big(\sum_{i=1}^d \beta_i \frac{\partial f^0_\theta}{\partial \theta^{(i)}} = 0 \implies \forall i = 1, \ldots, d,\ \beta_i = 0\ a.s\big)$ *or* $\big(\sum_{i=1}^d \beta_i \frac{\partial H^0_\theta}{\partial \theta^{(i)}} = 0 \implies \forall i = 1, \ldots, d,\ \beta_i = 0\ a.s\big)$.

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

Assumption *K*(Θ): *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 $\widetilde{\boldsymbol{K}}$ (Θ): *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 [34] and are not too restrictive: for instance, if the Lipschitz 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: for $0 \leq a < 1$
$$\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\partial_\theta f_\theta,\Theta)+\alpha_j(\partial_\theta M_\theta,\Theta)=O(a^j).$$
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: for $\gamma > 1$,

$$\alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta) = O(j^{-\gamma}).$$

- if $r \geq 3$ then
  - if $\gamma > 2$ then any (κ_n) 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 $\gamma > 1 + 3/r$ then any (κ_n) such as $1/\kappa_n = o(1)$ can be chosen;
  - if $1 < \gamma < 1 + 3/r$ then any (κ_n) such as $\kappa_n = n^\delta$ where $\delta > 1 + 3/r - \gamma$ 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. Also, κ_n can be chosen from a data-driven procedure; see Section 6 where the slope heuristic procedure is performed for the calibration of the penalty term.

3.2. 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*

$$\mathbb{P}(\widehat{m} = m^*) \underset{n\to\infty}{\longrightarrow} 1 \quad \textit{and} \quad \widehat{\theta}(\widehat{m}) \underset{n\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} \theta^*. \tag{3.2}$$

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 Var(Θ) holds, then*

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

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

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 (2.6) 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 (for instance 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(∞) 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:

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

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 (2.6) 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 $Id(\Theta)$, $\mathrm{Var}(\Theta)$ hold and that the sequence (ψ_k) is twice differentiable (with respect to θ) on Θ, with $\sum_k \|\partial_\theta^2 \psi_k(\theta)\|_\Theta < \infty$ and for $\gamma > 1$,

$$\|\psi_k(\theta)\|_\Theta + \|\partial_\theta \psi_k(\theta)\|_\Theta = O(k^{-\gamma}).$$

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 [16], $(X_t)_{t\in\mathbb{Z}}$ is an APARCH(δ,p,q) process with $p, q \geq 0$ if:

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

for any $t \in \mathbb{Z}$, 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 [7], with $\theta =$

$(\omega, \alpha_1, \dots, \alpha_p, \gamma_1, \dots, \gamma_p, \beta_1, \dots, \beta_p)'$, 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 $M_\theta^t = \sigma_t$, then $\alpha_k(M_\theta, \Theta) = \max(\|b_k^+(\theta)\|_\Theta^{1/\delta}, \|b_k^-(\theta)\|_\Theta^{1/\delta})$, 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

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

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 [16] and [40], we define $(X_t)_{t \in \mathbb{Z}}$ as an ARMA(p, q)-GARCH(p', q') process with $p, q, p', q' \geq 0$ if:

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

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

$$\Theta_{p,q,p',q'} = \Big\{ (a_1, \dots, d_{q'}) \in \mathbb{R}^{p+q+p'+1+q'}, \ \sum_{j=1}^{q'} d_j + \|\xi_0\|_4 \sum_{j=1}^{p'} c_j < 1$$

$$\text{and } \Big(1-\sum_{j=1}^{p} a_j z^j\Big)\,\Big(1-\sum_{j=1}^{q} b_j z^j\Big) \neq 0 \text{ for all } |z| \leq 1\Big\}.$$

Therefore, if $(a_1,\dots,d_{q'}) \in \Theta_{p,q,p',q'}$, $(\varepsilon_t)_t$ 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 [7], 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 that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA(p^*,q^*)-GARCH$(p^{'*},q^{'*})$ process with unknown orders $0 \leq p^* \leq p_{\max}$, $0 \leq q^* \leq q_{\max}$, $0 \leq p^{'*} \leq p'_{\max}$ and $0 \leq q^{'*} \leq q'_{\max}$ and unknown parameters: $c^*_0,\dots,c^*_{p^{'*}},d^*_1,\dots,d^*_{q^{'*}},a^*_1,\dots,a^*_{p^*},b^*_1,\dots,b_{q^*}$.

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

$$\begin{aligned}\theta^* = \big(c^*_0,\dots,c^*_{p^{'*}},0,\dots,0,d^*_1,\dots,d^*_{q^{'*}},0,\dots,0 \\ ,a^*_1,\dots,a^*_{p^*},0,\dots,0,b^*_1,\dots,b_{q^*},0,\dots,0\big) \in \mathbb{R}^d.\end{aligned}$$

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 [38], [39]) and the absolute residual (or usual residuals) correlogram (see for instance [37], [18], [36]), 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 follow the same scheme of procedure used in ([38], [39]) while relying on some of their results. But three main differences need to be pointed out:

- the results of Li and Mak (1994) are based on the exact likelihood of the data, which is then assumed to be known. But it is not at all the case even for simple ARMA$(1,1)$ or GARCH$(1,1)$ processes. By working directly on the quasi-likelihood, we really proposes a feasible Portemanteau test;
- we provide more detailed sufficient conditions to get the asymptotic results of the Portmanteau test;
- our procedure is also applied to the selected model, which is not necessarily the true model.

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

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

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

$$\widehat{\gamma}_k(m) := \frac{1}{n} \sum_{t=k+1}^{n} \big(\widehat{e}_t^2(m) - 1\big)\big(\widehat{e}_{t-k}^2(m) - 1\big)$$

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

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, with also*
- $\mathbb{E}[\xi_0^3] = 0$*;*
- $\displaystyle\sum_{t=1}^{\infty} t^{-1/4}\Big(\sum_{j \geq t} \alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta)\Big)^{1/2} < \infty$ *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 (7.38), 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:

$$\begin{cases} H_0 : \exists m^* \in \mathcal{M}, \text{ such as } (X_1, \dots, X_n) \text{ is a trajectory of } X \in \mathcal{AC}(M_\theta, f_{\theta^*}) \\[2ex] H_1 : \nexists m^* \in \mathcal{M}, \text{ such as } (X_1, \dots, X_n) \text{ is a trajectory of } X \in \mathcal{AC}(M_\theta, f_{\theta^*}) \end{cases}.$$

with $\theta^* \in \Theta(m^*)$ in both 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. In practice the constant μ_4 and the columns of the matrix $J_K(m^*)$ (see (7.34)) involved in $V(\theta^*, m^*)$ 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 $\big(\widehat{J}_K(\widehat{\theta}(\widehat{m})\big)_{\cdot,k} = \frac{1}{n}\sum_{t=1}^{n-k}[(\widehat{e}_t(\widehat{m}))^2 - 1]\partial_\theta \log\big(M^{t+k}_{\hat{\theta}(\widehat{m})}\big)$.
2. For AR(∞) models (and then for causal invertible ARMA(p,q)), since $M_\theta = \sigma$ as we have seen in Sub-section 4.1, we deduce from (7.38) that $V(\theta^*, m^*) = I_K$ as $J_K(m^*) = 0$. Hence, in such a case, we simply obtained:
$$\widehat{Q}_K(\widehat{m}) = n\left\|\widehat{\rho}(\widehat{m})\right\|^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K). \tag{5.3}$$
Note that working with autocorrelations of squared residuals rather than those of residuals, avoids the need to subtract the number of estimated parameters in the asymptotic chi-square distribution. Hence our result is valid for any $K \in \mathbb{N}^*$.

6. Numerical results

This section features some simulation experiments that are performed to assess the usefulness of the asymptotic results obtained in Section 3. 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 $\{100, 500, 1000, 2000\}$. The process used to generate the trajectory is indicated each time. Throughout this section, (ξ_t) represents a Gaussian white noise with variance unity.Various configurations studied are presented and we compare the performance of the penalties $\log n$ and $\sqrt{n}$ as well as a data-driven procedure based on the slope heuristic. This procedure (developed in [13],[14] and [6]) has been successfully applied to solve model selection questions in several situations (see for instance [8], [5], [10], [35]). Let us give a brief description of the calibration of κ_n by the slope heuristic.

Slope Heuristic

Leaving aside the theoretical details, the slope procedure is based on the fact that for "large" models, we expect that the quasi-log-likelihood $\widehat{L}_n\big(\widehat{\theta}(m)\big)$ linearly increases with the dimension $|m|$ when the family of models is hierarchical. Then, considering twice the slope of this linear part (say $\widehat{\kappa}_n$) allows for adaptive calibration of the penalty κ_n. This is done in two steps:

1. For each $0 \le |m| \le d$, consider the two-dimensional sample points $\big(|m|, \widehat{L}_n\big(\widehat{\theta}(m)\big)\big)$ where m is the best model (in term of maximum quasi-likelihood) of dimension $|m|$ and draw its graph;

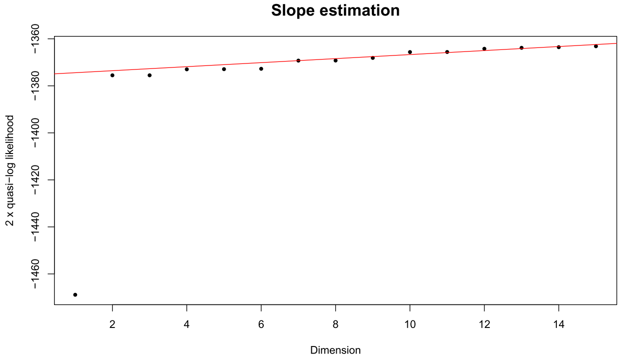


Fig 1. *The curve of quasi-log likelihood of an AR(2) versus the dimension; the oblique solid line with a slope* $\widehat{a} = 0.86$*, represents the linear part of the curve.*

2. Compute a least squares estimator $\widehat{a}$ of the slope of the right-side linear part of this curve, which can be selected from a change point detection (see for instance Figure 1 where we have plotted the quasi-log likelihood of Model 1 (defined in the subsection 6.1) with $n = 500$ and the candidates models are the $\mathrm{AR}(p), 0 \leq p \leq 15$);
3. Use $\widehat{\kappa}_n = 2\,\widehat{a}$ in the selection procedure (2.4).

Let us point out that, the theoretical validity of this procedure from the asymptotic point of view (efficiency) has not yet been widely studied. However, a theoretical validity from the non-asymptotic point of view has been studied in several setting. The classical measure used in this framework is the oracle inequality. For $\theta \in \Theta$, define the Kullback-Leibler divergence between the conditional density indexed by θ and the true one by

$$D_{KL}\big(\theta^* || \theta\big) := \frac{1}{2}\,\mathbb{E}[q_0(\theta)] - \frac{1}{2}\,\mathbb{E}[q_0(\theta^*)] \geq 0,$$

where the expectation is taken indeed under the distribution indexed by θ^*. The "ideal" model $m(\theta^*)$ (the one whose is closest to m^* according to the Kullback-Leibler risk) satisfying:

$$m(\theta^*) = \operatorname*{argmin}_{m \in \mathcal{M}} \mathbb{E}\big[D_{KL}\big(\theta^* || \widehat{\theta}_m\big)\big].$$

The model $m(\theta^*)$, which depends on the true distribution of the observations is called *the oracle* and cannot be computed in practice. The aim is to calibrate the penalty term, such that the chosen model $\widehat{m}$ provides a risk which is close as possible to the risk of the oracle; that is for instance

$$\mathbb{E}\big[D_{KL}\big(\theta^* || \widehat{\theta}_{\widehat{m}}\big)\big] \leq C\;\mathbb{E}\big[D_{KL}\big(\theta^* || \widehat{\theta}_{m(\theta^*)}\big)\big] + r_n$$

where C is a non-negative constant, expected to be close to 1 and (r_n) a sequence satisfying $n\,r_n = o(1)$. Such property has been established for the slope heuristic procedure in the Gaussian model selection, the penalized procedure for least-squares regression, the model selection for density estimation (see [14], [6], [35]) among others. Nevertheless, the theoretical validity for the class of models considered here has not yet been addressed either from asymptotic or non-asymptotic point of view. This issue could be an interesting extension of this work.

6.1. Monte-Carlo experiments for common time series selection

We first generate some classical models as "true" models m^*:

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: $X_t = \xi_t\sqrt{0.2 + 0.4X_{t-1}^2 + 0.2X_{t-2}^2}$;
4. Model 4, GARCH$(1,1)$ process: $\left\{\begin{array}{lcl} X_t & = & \sigma_t\,\xi_t \\ \sigma_t^2 & = & 0.2 + 0.3X_{t-1}^2 + 0.5\sigma_{t-1}^2 \end{array}\right.$.

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

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

As a consequence, there are 66 candidate models. Note also that in our simulations, since we have more than one model per dimension, slope estimation is done after considering the "best model" (which maximizes quasi-log likelihood) within each dimension.

The results of the model selection procedure are displayed in Table 1. More precisely, for each penalty ($\log n$, $\sqrt{n}$ and $\widehat{\kappa}_n$) the frequency that the associated criterion selects respectively a wrong model, the true model and an overfitted model (here a model that contains the true model).

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 non adaptive penalties $(\log n, \sqrt{n})$ lead to consistent criteria for the four models under consideration. Note also that the typical BIC $\log n$ penalty is more 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 is, compared to the $\log n$ penalty. One cannot also fail to mention that the slope heuristic is relatively better than the $\log n$ and $\sqrt{n}$ penalties for small samples but also asymptotically especially for GARCH type models. Let us recall that the theoretical validity of slope heuristic for the class of models considered here has not yet been established. These satisfactory results can be a motivation for investigating this issue.

In addition, 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

Table 1

Percentage of selected order based on 1000 independent replications depending on sample's length for the penalty $\log n$*,* $\sqrt{n}$ *and* $\widehat{\kappa}_n$*; where M1,$\cdots$, M4 refers to Model 1, 2, 3 and 4 respectively and W, T, O refers to wrong, true and overfitted selection.*

	n	100			500			1000			2000		
		$\log n$	$\sqrt{n}$	$\widehat{\kappa}_n$	$\log n$	$\sqrt{n}$	$\widehat{\kappa}_n$	$\log n$	$\sqrt{n}$	$\widehat{\kappa}_n$	$\log n$	$\sqrt{n}$	$\widehat{\kappa}_n$
M1	W	21.4	32.3	18.4	1.7	0.8	0.9	0.8	0.1	0.1	0.2	0	0
	T	74.2	67.6	79.7	97.2	99.2	99.1	98.2	99.9	99.9	99.2	100	100
	O	4.4	0.1	1.9	1.1	0	0	1.0	0	0	0.6	0	0
M2	W	30.4	57.7	28.0	4.8	4.2	4.0	0.7	0.3	0.3	0.4	0	0
	T	64.1	42.1	67.3	93.6	95.8	95.8	98.2	99.7	99.6	99.2	100	100
	O	5.5	0.2	4.7	1.6	0	0.2	1.1	0	0.1	0.4	0	0
M3	W	76.1	90.8	53.5	27.3	67.1	18.0	14.0	41.5	13.3	4.6	12.0	4.6
	T	23.8	9.2	39.8	72.7	32.9	79.9	85.9	58.5	86.7	95.4	88.0	95.4
	O	0.1	0	6.7	0	0	2.1	0.1	0	0	0	0	0
M4	W	83.8	94.3	73.4	22.1	61.5	20.4	5.8	31.3	5.7	1.8	6.2	0.7
	T	15.9	5.7	21.6	77.5	38.5	75.9	93.2	68.7	92.6	98.0	93.8	99.3
	O	0.3	0	5.0	0.4	0	3.7	1.0	0	1.7	0.2	0	0

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 and 4.

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 [38] suggested that the autocorrelations $\widehat{\rho}_k(\widehat{m})$ with $1 \leq k \leq K$ 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, [54] 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. 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 previously considered criteria on a particular case of dimension selection. The process generated data is considered as follows:

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 $K = 6$.

n		100		500		1000		2000	
		size	power	size	power	size	power	size	power
$K = 3$	Model 1	3.3	10.9	6.2	52.2	3.5	84.8	5.0	98.2
	Model 2	3.3	7.0	4.8	23.3	6.2	42.4	4.9	70.4
	Model 3	4.6	6.4	8.4	44.1	14.3	81.0	36.9	99.4
	Model 4	9.5	23.2	21.3	38.5	33.6	57.2	39.4	88.3
$K = 6$	Model 1	2.9	9.1	4.9	42.0	4.4	76.3	4.5	97.6
	Model 2	3.0	6.3	5.2	18.0	5.1	35.1	4.6	60.2
	Model 3	4.5	12.6	11.1	64.4	14.7	92.5	27.9	99.9
	Model 4	4.3	52.7	4.2	98.6	3.2	99.6	3.6	99.9

Table 3
Percentage of selected model based on 1000 replications depending on sample's length for Model 5

n	100			500			1000			2000		
	$\log n$	$\sqrt{n}$	$\widehat{\kappa_n}$	$\log n$	$\sqrt{n}$	$\widehat{\kappa_n}$	$\log n$	$\sqrt{n}$	$\widehat{\kappa_n}$	$\log n$	$\sqrt{n}$	$\widehat{\kappa_n}$
T	70.4	67.3	71.0	90	100	100	93.2	100	100	95.3	100	100
O	25.0	1.6	28.8	10	0	0	6.8	0	0	4.7	0	0
W	4.6	31.1	0.2	0	0	0	0	0	0	0	0	0

$$\text{Model 5} : X_t = 0.4X_{t-3} + 0.4X_{t-4} + \xi_t.$$

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

$$\mathcal{M} = \mathcal{P}(\{1, 2, \cdots, 10\})$$

$$\Longrightarrow X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots + \theta_{10} X_{t-10} + \xi_t$$
$$\text{and } \theta = (\theta_1, \theta_2, \cdots, \theta_{10})' \in \Theta(m).$$

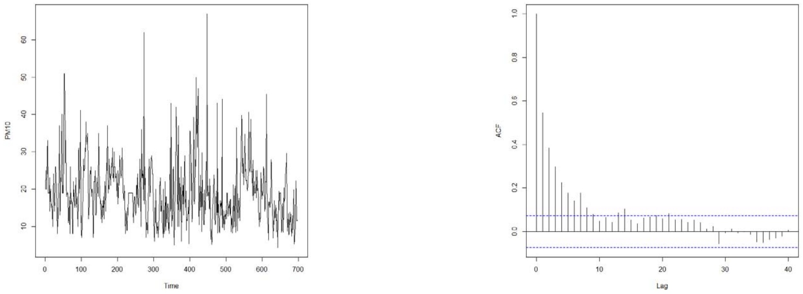
As a consequence, $1024 = 2^{10}$ candidate models are considered and Table 3 presents the results of the selection procedure.

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. Application to real data

6.3.1. Air quality analysis

Air quality, which can be defined as the level of cleanliness of the air, is probably one of the first health and environmental concerns of this new century. With



(a) Time plot of PM10 levels.

(b) ACF for PM10 with the bounds $\pm 1.96/\sqrt{n}$.

FIG 2. *The Marseille PM10 levels (January 1st, 2018 to November 30, 2019).*

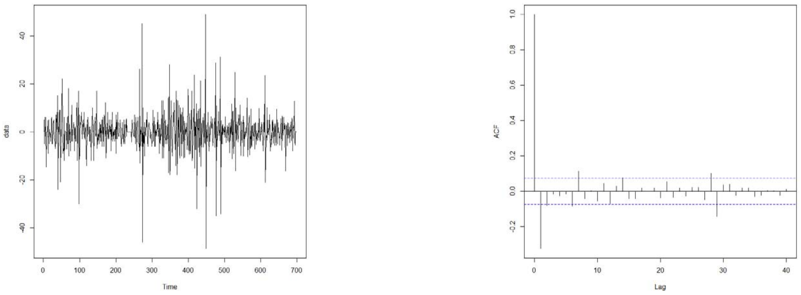
TABLE 4
Summary of the results of the model selection and goodness-of-fit analysis on PM10.

	$\kappa_n = \log(n)$	$\kappa_n = \sqrt{n}$	$\kappa_n = \widehat{\kappa}_n$
$\widehat{m}$	ARMA$(1,2)$	ARMA$(1,1)$	ARMA$(1,1)$
$\widehat{Q}_{10}(\widehat{m})$	11.09	18.02	18.02
$p-value$	0.35	0.055	0.055

the increasing number of human activities, the air is being degraded by a wide variety of pollutants, including PM. PM stands for particulate matter [22]: the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Let consider daily observations of PM10 (downloaded from Air PACA) at Marseille Kaddouz station (France) from January 1, 2018 to November 30, 2019. This is a time series trajectory of length $n = 698$ (see Figure 2a). We are going to use our model selection criteria to identify the "best" model for this time series.

An inspection of the Figure 2 may suggest us a family of candidate models. Fist, the slow decrease of the sample autocorrelation (up to lag 6), suggests that there is a component trend in the variability of the PM10. Also, a close inspection of the data shows that pollution is on average much lower on weekends than on working days. So before identifying a plausible family of models, let consider the detrended time series by differencing (see Figure 3). Therefore, we use the same family $\mathcal{M}$ already considered in Subsection 6.1 that provides us 66 candidate models. For each model, we compute the criterion (2.4) with $\kappa_n = \log(n)$, $\kappa_n = \sqrt{n}$ and also using an adaptive penalty $\widehat{\kappa}_n$ obtained from the slope heuristic procedure. The selection results and also the goodness-of-fit of the selected model are featured in the Table 4.

This table shows that all p-values are greater than 0.05, and then none of the test statistics leads us to reject the null hypothesis at this level even though the case of the ARMA$(1,1)$ is somehow limit. The chosen ARMA$(1,2)$ seems to be



(a) PM10 residuals levels.

(b) ACF for PM10 residuals with the bounds $\pm 1.96/\sqrt{n}$.

Fig 3. *Elimination of trend and seasonality in Marseille PM10 levels (January 1st, 2018 to November 30, 2019).*

Table 5

Summary of the results of the model selection and goodness-of-fit analysis on FTSE index.

	$\kappa_n = \log(n)$	$\kappa_n = \sqrt{n}$	$\kappa_n = \widehat{\kappa}_n$
$\widehat{m}$	GARCH$(1,1)$	GARCH$(1,1)$	GARCH$(1,1)$
$\widehat{Q}_{10}(\widehat{m})$	9.30	9.30	9.30
$p - value$	0.50	0.50	0.50

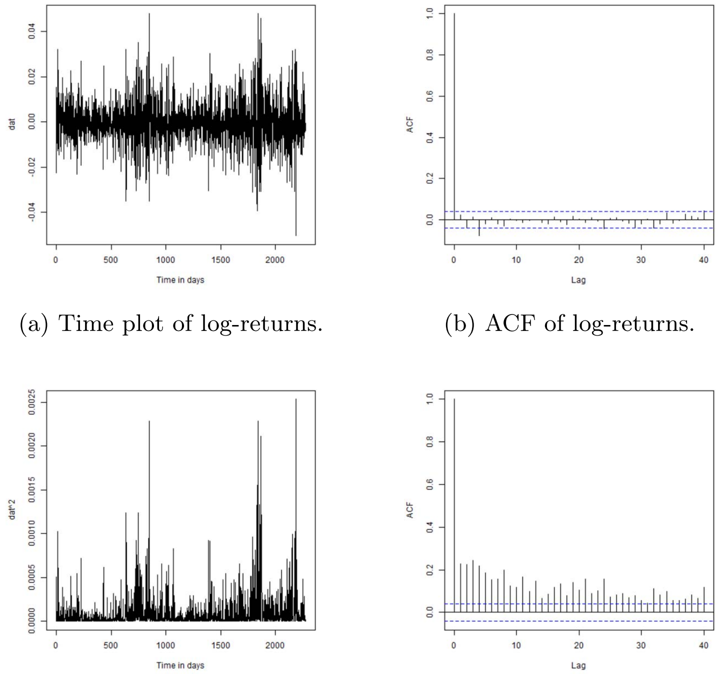
the more suitable model for PM10 time series.

6.3.2. Financial index analysis

We consider the returns of the daily closing prices of the FTSE 100 index and also the SP 500. They are respectively 2273 and 2264 observations from January 4th, 2010 to December 31st, 2018 for FTSE 100 and SP500. The time plot and the correlograms for the log-returns and squared log-returns are plotted in Figure 4. Figures 4a and 4c exhibit the conditional heteroskedasticity in the log-return time series. Moreover, Figure 4b 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 4d suggests that the strong white noise assumption cannot be sustained for this log-returns sequence of FTSE index. We also have the same conclusion for SP 500 (see Figure 5)

As in the previous illustrative example, the ARMA-GARCH is a plausible family for modeling of the FTSE 100 and SP 500 index. The penalization $\log n$, $\sqrt{n}$ and $\widehat{\kappa}_n$ have been applied to identify the best order and the goodness-of-fit of the selected model has been tested by the Portmanteau test.

The GARCH$(1,1)$ is the "best" model based on the three criteria considered and it is adequate (at level 0.95) to model the FTSE 100 index. Regarding the SP 500 index, the GARCH$(1,1)$ is still the best model based on all three criteria and $\widehat{Q}_{10}(\widehat{m}) = 15.2$ associated with a p-value of 0.12. These results



(a) Time plot of log-returns.

(b) ACF of log-returns.

(c) Time plot of squared log-returns.

(d) ACF of squared log-returns.

Fig 4. *Daily closing FTSE 100 index (January 4th, 2010 to December 31 st, 2018).*

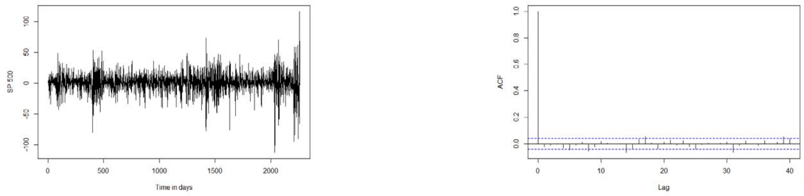
are not surprising since the GARCH$(1,1)$ is the reference model and the most commonly used in empirical studies. In addition, [20] found the GARCH$(1,1)$ to be adequate using a FTSE 100 trajectory from April 3, 1984 to April 3, 2007 and January 3, 1950 to July 24, 2009 for SP 500.

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 $0 \leq \delta \leq \min(d_1, d_2)$ 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 $0 \leq \delta' \leq \min(d_1, d_2)$ 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 $m^{(i)}_0 = 0$.



(a) Time plot of log-returns .

(b) ACF with the bounds ±$1.96/\sqrt{n}$.

Fig 5. *Daily closing price of SP500 (January 4th, 2010 to December 31 st, 2018).*

Consider now $\theta = {}^t(\alpha, \alpha'_1, \alpha'_2, \beta, \beta'_1, \beta'_2) \in \mathbb{R}^d$ (and therefore $\max(d_1, d_2) \leq d \leq d_1 + d_2$), $f_\theta = h_\alpha + \ell^{(1)}_{\alpha'_1} + \ell^{(2)}_{\alpha'_2}$ 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 $t \in \mathbb{Z}$,

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

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

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 [34], with $r \leq 3$, (7.1) is established when:

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

With $r \geq 3$, and under the assumptions, we first recall some results already obtained in [9]: for any $t \in \mathbb{Z}$,

- $\mathbb{E}\big[|X_t|^r + \|f^t_\theta\|^r_\Theta + \|\widehat{f}^t_\theta\|^r_\Theta + \|M^t_\theta\|^r_\Theta + \|\widehat{M}^t_\theta\|^r_\Theta + \|H^t_\theta\|^{r/2}_\Theta + \|\widehat{H}^t_\theta\|^{r/2}_\Theta\big] < \infty$(7.3)

- $$\begin{cases} \mathbb{E}\big[\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^r\big] \le C\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)\Big)^r \\ \mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r\big] \le C\Big(\sum_{j\ge t}\alpha_j(M_\theta,\Theta)\Big)^r \\ \mathbb{E}\big[\|H_\theta^t - \widehat{H}_\theta^t\|_\Theta^{r/2}\big] \le C\Big(\min\Big\{\sum_{j\ge t}\alpha_j(M_\theta,\Theta)\ ,\ \sum_{j\ge t}\alpha_j(H_\theta,\Theta)\Big\}\Big)^{r/2}. \end{cases} \tag{7.4}$$

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

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

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

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

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

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

Again with Hölder's inequality and (7.3),

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

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,

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

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 [34], with $r \leq 4$, (7.1) is established when:

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

By proceeding as in the previous case, we deduce

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

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

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

■

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 $r \geq 2$. 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:

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

Using again Corollary 1 of [34], it is sufficient to prove for $r \leq 3$ that

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

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,

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

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

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

Thus,

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

Using again the results of [9], we know that:

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

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

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

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

Using (7.13) and (7.14), we deduce

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

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 $X\subset\widetilde{\mathcal{A}C}(\widetilde{H}_\theta)$, 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,

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

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 [34]. ■

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:*

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

Proof. We have:

$$\begin{aligned}\frac{1}{\kappa_n} \left| \widehat{L}_n(\widehat{\theta}(m)) - \widehat{L}_n\big(\widehat{\theta}(m^*)\big) \right| &= \frac{1}{\kappa_n} \Big| \widehat{L}_n(\widehat{\theta}(m)) - L_n(\widehat{\theta}(m)) + L_n(\widehat{\theta}(m)) - L_n\big(\widehat{\theta}(m^*)\big) \\ &\qquad + L_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m^*)\big) \Big| \\ &\leq \frac{2}{\kappa_n} \big\| \widehat{L}_n(\theta) - L_n(\theta) \big\|_\Theta + \frac{1}{\kappa_n} \big| L_n(\widehat{\theta}(m)) - L_n\big(\widehat{\theta}(m^*)\big) \big|.\end{aligned}$$

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

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

Since

$$\frac{1}{\kappa_n} \Big| L_n(\widehat{\theta}(m)) - L_n\big(\widehat{\theta}(m^*)\big) \Big| \leq \frac{1}{\kappa_n} \Big| L_n(\widehat{\theta}(m)) - L_n(\theta^*) \Big| + \frac{1}{\kappa_n} \Big| L_n(\widehat{\theta}(m^*)) - L_n(\theta^*) \Big|.$$

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:

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

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

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

Since $\frac{1}{\kappa_n} = o(1)$ and from [9] then $\sqrt{n}\big(\widehat{\theta}(m) - \theta^*\big) = O_P(1)$ and $\frac{1}{\sqrt{n}} \frac{\partial L_n(\widehat{\theta}(m))}{\partial \theta} = o_P(1)$, it follows that:

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

On the other hand, for the second term of the right hand side of equality (7.17), let us note that, we have from [9]:

- $\sqrt{n}\,\big(\widehat{\theta}(m)-\theta^*\big)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{A}_{\theta^*,m}$ a Gaussian random variable from Theorem 2 of [9].
- $-\dfrac{2}{n}\Big(\dfrac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\Big)_{i,j\in m}\xrightarrow[n\to+\infty]{a.s.}F(\theta^*,m)$ since $\widehat{\theta}(m)\xrightarrow[n\to+\infty]{a.s.}\theta^*$ and using the assumption $\mathrm{Var}(\Theta)$ insuring that the matrix $F(\theta^*,m)$ exists and is definite positive (see also [9]).

Hence,

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

We deduce that

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

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

7.1. Misspecified model

When a model m is misspecified ($\theta^*\notin\Theta(m)$), we will show that $\mathbb{P}(m^*\not\subseteq\widehat{m})\xrightarrow[n\to\infty]{}0$ by following the key idea of similar proof in [49] by defining the "best" parameter $\theta^*(m)\in\Theta(m)$ which will play the role of θ^* in cases of "true" or overfitted model. For model $m\in\mathcal{M}$, let define

$$\theta^*(m):=\operatorname*{argmax}_{\theta\in\Theta(m)}L(\theta)\quad\text{with}\quad L(\theta):=-\frac{1}{2}\,\mathbb{E}[q_0(\theta)].\tag{7.20}$$

Proposition 2. *For any model $m\in\mathcal{M}$, there exists $\theta^*(m)$ in $\Theta(m)$. Moreover, under the Identifiability assumption $Id(\Theta(m))$, $\theta^*(m)$ is unique.*

Proof. Let recall from the Subsection 2.1 when deriving the Gaussian conditional likelihood that $q_t(\theta)$ is none other than -2 times the conditional Gaussian log-density (with mean f^t_θ and variance H^t_θ) at the observation X_t. Next, let define the Kullback Leiber divergence between the conditional density indexed by θ and the true one indexed by θ^*,

$$D_{KL}\big(\theta^*||\theta\big):=\mathbb{E}\Big[\log\Big(\frac{\exp\big(-0.5\times q_t(\theta^*)\big)}{\exp\big(-0.5\times q_t(\theta)\big)}\Big)\Big]=-\frac{1}{2}\,\mathbb{E}[q_0(\theta^*)]+\frac{1}{2}\,\mathbb{E}[q_0(\theta)],$$

where the expectation is taken indeed under the distribution indexed by θ^*.

Moreover, since minimizing the Kullback discrepency over $\Theta(m)$ is equivalent to maximize $L(\theta)$,

$$\underset{\theta\in\Theta(m)}{\text{argmin}}\, D_{KL}\big(\theta^*||\theta\big) = \underset{\theta\in\Theta(m)}{\text{argmax}}\, L(\theta) = \theta^*(m),$$

it follows that $\theta^*(m)$ is the Kullback Leiber projection of the true density distribution onto the set of distributions generated by $\Theta(m)$, which ends the proof of existence.

On the other hand the uniqueness is a consequence of $Id(\Theta(m))$. Indeed, since $\theta^*(m) \in \Theta(m)$, there is no other parameter $\theta_0 \in \Theta(m)$ such that almost surely, we have

$$\Big(f^0_{\theta^*(m)} = f^0_{\theta_0} \quad \text{and} \quad M^0_{\theta^*(m)} = M^0_{\theta_0}\Big) \quad \text{which implies} \quad L\big(\theta^*(m)\big) = L(\theta_0). \quad \blacksquare$$

It is worth noting, since $L(\theta)$ has a unique maximum in θ^* (see [9]), and along with the fact that $\theta^* \in \Theta(m)$, it follows that $\theta^*(m) = \theta^*$ when m is the true model or an overfitted one.

Let us show that even in the presence of misspecification, the QMLE still remains consistent but for $\theta^*(m)$. This important result will allow us to show that our model selection procedure can not choose a misspecified model.

Proposition 3. *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 assumptions $Id(\Theta(m))$, $D\big(\Theta(m)\big)$ and $K\big(\Theta(m)\big)$ hold for a compact set $\Theta(m) \subset \Theta$, it holds*

$$\Big\|\frac{1}{n}\, L_n(\theta) - L(\theta)\Big\|_{\Theta(m)} \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} 0 \quad \text{and} \tag{7.21}$$

$$\widehat{\theta}(m) \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} \theta^*(m). \tag{7.22}$$

Proof. The proof of (7.21) follows from a consequence of uniform strong law of large numbers for stationary ergodic sequence (see the proof of Theorem 1 in [9]). The second result holds by applying (7.21) and Lemma 1. ■

7.2. Proof of Theorem 3.1

Before diving into the proof, remark first that:

$$\mathbb{P}(\widehat{m} = m^*) = 1 - \mathbb{P}(m^* \subset \widehat{m}) - \mathbb{P}(m^* \not\subseteq \widehat{m}). \tag{7.23}$$

As we point out in Subsection 2.1, the proof is divided into two parts; the first part shows that our selection criterion choses an overfitted model with probability decreasing to zero while the second part shows a similar behavior for the probability of selecting a misspecified model.

Proof. 1. Since $\mathcal{M}$ is finite, let $m_0 \in \mathcal{M}$ such as $\widehat{m} = m_0$ and $m^* \subset m_0$, (i.e an overfitted model was selected, but let show that this cannot happen). Let compute $\mathbb{P}\big(\widehat{C}(m_0) \leq \widehat{C}(m^*)\big)$ for large n.

We have:

$$\begin{aligned}
\mathbb{P}\big(\widehat{C}(m_0) \leq \widehat{C}(m^*)\big) &= \mathbb{P}\Big(-2\,\widehat{L}_n\big(\widehat{\theta}(m_0)\big) + |m_0|\,\kappa_n \leq -2\,\widehat{L}_n\big(\widehat{\theta}(m^*)\big) + |m^*|\,\kappa_n\Big) \\
&= \mathbb{P}\Big(-2\,\widehat{L}_n\big(\widehat{\theta}(m_0)\big) + 2\,\widehat{L}_n\big(\widehat{\theta}(m^*)\big) \leq \kappa_n(|m^*| - |m_0|)\Big) \\
&= \mathbb{P}\Big(\frac{1}{\kappa_n}\big(\widehat{L}_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m_0)\big) \leq \frac{(|m^*| - |m_0|)}{2}\Big) \\
&\underset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

by vertue of Lemma 3 and because $|m_0| - |m^*| \geq 1$.

This shows, $\widehat{C}(m_0) > \widehat{C}(m^*)$ with probability going to 1, i.e. $\widehat{C}(\widehat{m}) > \widehat{C}(m^*)$. We get a contradiction along with definition of $\widehat{m}$ (2.4), and then the selection criteria can not choose $\widehat{m}$ which stricly contains the true model, thus

$$\mathbb{P}(m^* \subset \widehat{m}) \underset{n\to\infty}{\longrightarrow} 0.$$

2. Since $\mathcal{M}$ is finite, let $m_0 \in \mathcal{M}$ such as $\widehat{m} = m_0$ and $m^* \not\subseteq \widehat{m}$. Let compute $n^{-1}\big[\widehat{C}(m_0) - \widehat{C}(m^*)\big]$ for large n. First,

$$\begin{aligned}
\frac{1}{n}\Big[\widehat{L}_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m_0)\big)\Big] &= \frac{1}{n}\Big[L_n\big(\widehat{\theta}(m^*)\big) - L_n\big(\widehat{\theta}(m_0)\big)\Big] + o_{a.s}(1) \text{ with Lemma 1} \\
&= L\big(\widehat{\theta}(m^*)\big) - L\big(\widehat{\theta}(m_0)\big) + o_{a.s}(1) \text{ using Proposition 3} \\
&= \big[L\big(\widehat{\theta}(m^*)\big) - L(\theta^*)\big] - \big[L\big(\widehat{\theta}(m_0)\big) - L(\theta^*(m_0))\big] \\
&\qquad + \big[L(\theta^*) - L(\theta^*(m_0))\big] + o_{a.s}(1).
\end{aligned}$$

Since L is continuous over Θ, using continuous mapping theorem and the relation (7.22) of Proposition 3, it holds for n large enough

$$L\big(\widehat{\theta}(m^*)\big) - L(\theta^*) = o_{a.s}(1) \quad \text{and} \quad L\big(\widehat{\theta}(m_0)\big) - L(\theta^*(m_0)) = o_{a.s}(1).$$

Hence,

$$\frac{1}{n}\Big[\widehat{L}_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m_0)\big)\Big] = D_{KL}\big(\theta^*||\theta^*(m_0)\big) + o_{a.s}(1). \tag{7.24}$$

Note also that $D_{KL}\big(\theta^*||\theta^*(m_0)\big) > 0$ since $\theta^* \notin \Theta(m)$. As a consequence,

$$\frac{\widehat{C}(m_0) - \widehat{C}(m^*)}{n} = D_{KL}\big(\theta^*||\theta^*(m_0)\big) + \frac{\kappa_n}{n}(|m_0| - |m^*|) + o_{a.s}(1). \tag{7.25}$$

Moreover, since $\kappa_n = o(n)$ and all the considered models are finite dimensional, the equality (7.25) implies for large n that $\widehat{C}(m_0) > \widehat{C}(m^*)$ almost surely.

This means that it was possible to select a model $\widehat{m}$ with $\widehat{C}(\widehat{m}) > \widehat{C}(m^*)$, which is impossible according to the definition (2.4). Therefore the event $m^* \not\subseteq \widehat{m}$ can not happen and then

$$\mathbb{P}(m^* \not\subseteq \widehat{m}) \underset{n\to\infty}{\longrightarrow} 0.$$

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:

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

From the strong consistency of the QMLE (see New version of Theorem 1 of [9]), 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 $x = (x_i)_{1\le i\le d} \in \mathbb{R}^d$, denote $F_n(x) = \mathbb{P}\Big(\bigcap_{1\le i\le d} \sqrt{n}\,\big(\widehat{\theta}(\widehat{m})-\theta^*\big)_i \le x_i\Big)$.

First, we have:

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

Under the assumptions of Theorem 3.1,

$$\mathbb{P}\big(\widehat{m} = m^*\big) \underset{n\to\infty}{\longrightarrow} 1 \text{ and } \mathbb{P}\big(\widehat{m} \neq m^*\big) \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore the second term in the right side of the previous equality asymptotically vanishes. For the first term, we can write,

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

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

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

with Z a standard Gaussian random vector in $\mathbb{R}^{|m^*|}$ from the central limit theorem in Theorem 2 of [9], and this achieves the proof of 3.3 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:

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

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

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

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:

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

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 $k\in\mathbb{N}$,

$$\sqrt{n}\,\big\|\widehat{\gamma}_k(\theta)-\gamma_k(\theta)\big\|_\Theta\;\underset{n\to\infty}{\overset{a.s.}{\longrightarrow}}\;0. \tag{7.26}$$

We have:

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

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

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

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,

$$\begin{aligned}\mathbb{E}\Big[\Big\|\big(\widehat{e}_{t-k}^2(\theta)-e_{t-k}^2(\theta)\big)\big(\widehat{e}_t^2(\theta)-e_t^2(\theta)\big)\Big\|_\Theta^{1/2}\Big] &\le \Big(\mathbb{E}\big[\big\|\widehat{e}_t^2(\theta)-e_t^2(\theta)\big\|_\Theta\big] \\ &\qquad \times\mathbb{E}\big[\big\|\widehat{e}_{t-k}^2(\theta)-e_{t-k}^2(\theta)\big\|_\Theta\big]\Big)^{1/2}.\end{aligned}$$

But we have

$$\begin{aligned}\big\|\widehat{e}_t^2(\theta)-e_t^2(\theta)\big\|_\Theta &\le \frac{1}{\underline{h}}\big(2|X_t|+\|\widehat{f}_\theta^t\|_\Theta+\|f_\theta^t\|_\Theta\big)\big\|\widehat{f}_\theta^t-f_\theta^t\|_\Theta \\ &\qquad + \frac{4}{\underline{h}^{3/2}}\big(|X_t|^2+\|f_\theta^t\|_\Theta^2\big)\big\|\widehat{M}_\theta^t-M_\theta^t\|_\Theta.\end{aligned}$$

Therefore,

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

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

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

Therefore, from [34], $\|I_1^1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ when

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

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

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

Since $\mathbb{E}\big[\|e_t^2(\theta)\|_\Theta\big]\le 2\,\underline{h}^{-1}\mathbb{E}\big[X_t^2+\|f_\theta^t\|_\Theta^2\big]<\infty$ and similarly $\mathbb{E}\big[\|\widehat{e}_t^2(\theta)\|_\Theta\big]<\infty$, 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

$$\sum_{t=1}^{\infty}t^{-1/4}\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\widetilde{H}_\theta,\Theta)\Big)^{1/2}<\infty, \tag{7.29}$$

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 (7.22).

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 (7.29) that

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

This also implies

$$\sqrt{n}\big|\widehat{\rho}_k(m^*)-\rho_k(m^*)\big| \xrightarrow[n\to+\infty]{a.s.} 0 \quad \text{for any } k\geq 1. \tag{7.31}$$

(ii) The proof of this result has already been done in [38] 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 $k=1,\dots,K$,

$$\sqrt{n}\,\gamma_k(m^*)=\sqrt{n}\,\gamma_k(\widehat{\theta}(m^*))=\sqrt{n}\,\gamma_k(\theta^*)+\partial_\theta\gamma_k(\overline{\theta}^{(k)})\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta_i^*\big)_{i\in m^*}, \tag{7.32}$$

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 radius $\|(\widehat{\theta}(m^*)-\theta^*)_{i\in m^*}\|$ and centre θ^*. We also have

$$\begin{aligned}\partial_\theta\gamma_k(\theta)=-\frac{2}{n}\Big(\sum_{t=k+1}^{n} e_t^2(\theta)\,\big(e_{t-k}^2(\theta)-1\big)\frac{\partial_\theta M_\theta^t}{M_\theta^t}+e_t(\theta)\big(e_{t-k}^2(\theta)-1\big)\,\frac{\partial_\theta f_\theta^t}{M_\theta^t}\\ +e_{t-k}(\theta)\,\big(e_t^2(\theta)-1\big)\frac{\partial_\theta f_\theta^{t-k}}{M_\theta^{t-k}}+e_{t-k}^2(\theta)\,\big(e_t^2(\theta)-1\big)\frac{\partial_\theta M_\theta^{t-k}}{M_\theta^{t-k}}\Big).\end{aligned} \tag{7.33}$$

But

$$\mathbb{E}\Big[e_{t-k}(\theta^*)\,\big(e_t^2(\theta^*)-1\big)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}}\mid\sigma\big((\xi_s)_{s\leq t-k}\big)\Big]=e_{t-k}(\theta^*)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}}\mathbb{E}\big[e_t^2(\theta^*)-1\big]=0$$

since we 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 (7.33) 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 $\partial_{\theta^2}^2 f_\theta$ and $\partial_{\theta^2}^2 M_\theta$ 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

$$\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_{1\leq k\leq K}\xrightarrow[n\to+\infty]{a.s.} J_K(m^*)=-2\,\Big(\mathbb{E}\Big[\big(\xi_0^2-1\big)\,\frac{\partial}{\partial\theta_j}\log\big(M_{\theta^*}^i\big)\Big]\Big)_{1\leq i\leq K,\,j\in m^*}. \tag{7.34}$$

Under the assumptions, a central limit theorem for $\widehat{\theta}(m^*)$ has been established in [9], and this implies

$$\big(\partial_\theta\gamma_k(\overline{\theta}^{(k)})\big)_{1\leq k\leq K}\sqrt{n}\,\big((\widehat{\theta}(m^*))_i-\theta_i^*\big)_{i\in m^*}$$

$$\xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\Big(0\,,\; J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J'_K(m^*)\Big). \quad (7.35)$$

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 $\mu_4 - 1$ with $\mu_4 = \mathbb{E}[\xi_0^4]$. In such as case, the asymptotic behavior of the covariograms is well known and we deduce:

$$\sqrt{n}\,\big(\gamma_k(\theta^*)\big)_{1\le k\le K} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\big(0\,,\,(\mu_4-1)^2\,I_K\big), \quad (7.36)$$

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

We would like to use (7.32) for obtaining the asymptotic behavior of $\gamma(m^*)$. In (7.35) and (7.36), 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 [9]), 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:

$$-\;J_K(m^*)\,F(\theta^*,m^*)^{-1}\mathbb{E}\Big[\frac{\partial}{\partial\theta}L_n(\theta^*)\,\gamma(\theta^*)'\Big]. \quad (7.37)$$

From (2.1), a direct differentiation of L_n provides

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

so that,

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

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. Thus

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

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

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

Moreover the vector $\gamma(m^*)$ is normal distributed from Lemma 3.3 of [39].

Thus, using Slutsky Lemma and with $\gamma_0(m^*)\ \xrightarrow[n\to+\infty]{a.s.}\ \mu_4-1$, and with $\rho_k(m^*)=\gamma_k(m^*)/\gamma_0(m^*)$, the limit theorem (5.1) holds with

$$\begin{aligned}V(\theta^*,m^*):=I_K+(\mu_4-1)^{-2}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J'_K(m^*)&\\ +\,(\mu_4-1)^{-1}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\,J'_K(m^*).&\qquad(7.38)\end{aligned}$$

The proof is achieved after using the limit theorem (7.31).

(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 $F_n(x)=\mathbb{P}\Big(\bigcap\limits_{1\leq k\leq K}\sqrt{n}\,\big(\widehat{\rho}(\widehat{m})\big)_k\leq x_k\Big)$ the distribution function of $\sqrt{n}\widehat{\rho}(\widehat{m})$. Applying the Total Probability Rule and by virtue of Theorem 3.1, we obtain:

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

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

Acknowledgements

The authors are grateful to the referee for many relevant suggestions and comments that helped to notably improve the contents of the paper.

References

[1] H Akaike, *Information theory and an extension of the maximum likelihood principle*, Proceedings of the 2nd international symposium on information, Akademiai Kiado, Budapest (1973). MR0483125

[2] D.M. Allen, *The relationship between variable selection and data agumentation and a method for prediction*, Technometrics **16** (1974), no. 1, 125–127.

[3] P. Alquier and O. Wintenberger, *Model selection for weakly dependent time series forecasting*, Bernoulli **18** (2012), no. 3, 883–913. MR2948906

[4] O. Arkoun, J.-Y. Brua, and S. Pergamenshchikov, *Sequential model selection method for nonparametric autoregression*, Sequential Anal. **38** (2019), no. 4, 437–460. MR4057153

[5] S. Arlot, *Minimal penalties and the slope heuristics: a survey*, Journal de la SFDS **160** (2019), 1–106. MR4021422

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

[7] J.-M. Bardet, Y. Boularouk, and K. Djaballah, *Asymptotic behavior of the laplacian quasi-maximum likelihood estimator of affine causal processes*, Electronic journal of statistics **11** (2017), no. 1, 452–479. MR3619313

[8] J.-M. Bardet, W.C. Kengne, and O. Wintenberger, *Detecting multiple change-points in general causal time series using penalized quasi-likelihood*, Electronic journal of statistics **6** (2012), 435–477. MR2988415

[9] J.-M. Bardet and O. Wintenberger, *Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes*, The Annals of Statistics **37** (2009), no. 5B, 2730–2759. MR2541445

[10] J.-P. Baudry, C. Maugis, and B. Michel, *Slope heuristics: overview and implementation*, Statistics and Computing **22** (2012), no. 2, 455–470. MR2865029

[11] EML Beale, MG Kendall, and DW Mann, *The discarding of variables in multivariate analysis*, Biometrika **54** (1967), no. 3-4, 357–366. MR221656

[12] I. Berkes, L. Horváth, and P. Kokoszka, *GARCH processes: structure and estimation*, Bernoulli **9** (2003), 201–227. MR1997027

[13] L. Birgé and P. Massart, *Gaussian model selection*, Journal of the European Mathematical Society **3** (2001), no. 3, 203–268. MR1848946

[14] L. Birgé and P. Massart, *Minimal penalties for gaussian model selection*, Probability theory and related fields **138** (2007), no. 1-2, 33–73. MR2288064

[15] J. Ding, V. Tarokh, and Y. Yang, *Model selection techniques: An overview*, IEEE Signal Processing Magazine **35** (2018), no. 6, 16–34.

[16] Z. Ding, C. Granger, and R.F. Engle, *A long memory property of stock market returns and a new model*, Journal of empirical finance **1** (1993), no. 1, 83–106.

[17] P. Doukhan and O. Wintenberger, *Weakly dependent chains with infinite memory*, Stochastic Processes and their Applications **118** (2008), no. 11, 1997–2013. MR2462284

[18] P. Duchesne and C. Francq, *On diagnostic checking time series models with*

portmanteau test statistics based on generalized inverses and, COMPSTAT 2008, Springer, 2008, pp. 143–154.

[19] C. Francq and J.-M. Zakoïan, *Maximum likelihood estimation of pure garch and arma-garch processes*, Bernoulli **10** (2004), 605–637. MR2076065

[20] C. Francq and J.-M. Zakoian, *Garch models: structure, statistical inference and financial applications*, John Wiley & Sons, 2010. MR3185978

[21] J. Gao and H. Tong, *Semiparametric non-linear time series model selection*, Journal of the Royal Statistical Society: Series B **66** (2004), no. 2, 321–336. MR2062379

[22] United States government, *Particulate matter (pm) pollution*, `http://www.epa.gov/pm-pollution/particulate-matter-pm-basics`, 2017.

[23] E.J. Hannan, *The estimation of the order of an arma process*, The Annals of Statistics **8** (1980), no. 5, 1071–1081. MR2192006

[24] Ronald R Hocking and RN Leslie, *Selection of the best subset in regression analysis*, Technometrics **9** (1967), no. 4, 531–540. MR219192

[25] A.E. Hoerl and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[26] H.-L. Hsu, C.-K. Ing, and H. Tong, *On model selection from a finite family of possibly misspecified time series models*, The Annals of Statistics **47** (2019), no. 2, 1061–1087. MR3909960

[27] C.M. Hurvich and C.-L. Tsai, *Regression and time series model selection in small samples*, Biometrika **76** (1989), no. 2, 297–307. MR1016020

[28] C.-K. Ing, *Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series*, The Annals of Statistics **35** (2007), no. 3, 1238–1277. MR2341705

[29] C.-K. Ing, C.-Y. Sin, and S.-H. Yu, *Model selection for integrated autoregressive processes of infinite order*, Journal of Multivariate Analysis **106** (2012), 57–71. MR2887680

[30] C.-K. Ing and C.-Z. Wei, *Order selection for same-realization predictions in autoregressive processes*, The Annals of Statistics **33** (2005), no. 5, 2423–2474. MR2211091

[31] T. Jeantheau, *Strong consistency of estimators for multivariate arch models.*, Econometric Theory **14** (1998), no. 1, 70–86. MR1613694

[32] G. Kapetanios, *Model selection in threshold models*, Journal of Time Series Analysis **22** (2001), no. 6, 733–754. MR1867396

[33] A.B. Kock, *Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions*, Econometric Theory **32** (2016), no. 1, 243–259. MR3442507

[34] E.G. Kounias and T. Weng, *An inequality and almost sure convergence*, The Annals of Mathematical Statistics **40** (1969), no. 3, 1091–1093. MR0245058

[35] M. Lerasle, *Optimal model selection for density estimation of stationary data under various mixing conditions*, The Annals of Statistics **39** (2011), no. 4, 1852–1877. MR2893855

[36] G. Li and W.K. Li, *Least absolute deviation estimation for fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity*, Biometrika **95** (2008), no. 2, 399–414. MR2521590

[37] W.K. Li, *On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling*, Biometrika **79** (1992), no. 2, 435–437. MR1185148

[38] W.K. Li and T.K. Mak, *On the squared residual autocorrelations in nonlinear time series with conditional heteroskedasticity*, Journal of Time Series Analysis **15** (1994), no. 6, 627–636. MR1312326

[39] S. Ling and W.-K. Li, *Diagnostic checking of nonlinear multivariate time series with multivariate arch errors*, Journal of Time Series Analysis **18** (1997), no. 5, 447–464. MR1482451

[40] S. Ling and M. McAleer, *Asymptotic theory for a vector arma-garch model*, Econometric theory **19** (2003), no. 2, 280–310. MR1966031

[41] C.L Mallows, *Some comments on cp*, Technometrics **15** (1973), no. 4, 661–675.

[42] A. McQuarrie and C.L. Tsai, *Regression and time series model selection*, World Scientific Pub Co Inc, 1998. MR1641582

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

[44] Y. Ren and X. Zhang, *Subset selection for vector autoregressive processes via adaptive lasso*, Statistics & probability letters **80** (2010), no. 23-24, 1705–1712. MR2734232

[45] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics **6** (1978), 461–464. MR468014

[46] Q. Shao and L. Yang, *Oracally efficient estimation and consistent model selection for auto-regressive moving average time series with trend*, Journal of the Royal Statistical Society: Series B **79** (2017), no. 2, 507–524. MR3611757

[47] P. Shi and C.-L. Tsai, *Regression model selection-a residual likelihood approach*, Journal of the Royal Statistical Society: Series B **64** (2002), no. 2, 237–252. MR2887680

[48] R. Shibata, *Asymptotically efficient selection of the order of the model for estimating parameters of a linear process*, The Annals of Statistics (1980), 147–164. MR557560

[49] C.-Y. Sin and H. White, *Information criteria for selecting possibly misspecified parametric models*, Journal of Econometrics **71** (1996), no. 1-2, 207–225. MR1381082

[50] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, Journal of the royal statistical society. Series B (1974), 111–147. MR0415910

[51] D. Straumann and T. Mikosch, *Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach*, The Annals of Statistics **34** (2006), no. 5, 2449–2495. MR2291507

[52] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (1996), 267–288. MR2815776

[53] R.S. Tsay, *Order selection in nonstationary autoregressive models*, The Annals of Statistics **12** (1984), no. 4, 1425–1433. MR760697

[54] Y.K. Tse and X.L. Zuo, *Testing for conditional heteroscedasticity: Some*

monte carlo results, Journal of Statistical Computation and Simulation **58** (1997), no. 3, 237–253.

[55] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica (1982), 1–25. MR0688736

[56] H. Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101** (2006), no. 476, 1418–1429. MR2279469

[57] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B **67** (2005), no. 2, 301–320. MR2210692