Model selection for affine causal processes

Kare Kamila

Université Paris 1 Panthéon-Sorbonne

École Doctorale Sciences Mathématiques de Paris Centre (ED 386)

Laboratoire : Statistique, Analyse et Modélisation Multidisciplinaire, EA 4543

Sélection de modèles pour les séries affines causales

KAMILA KARE

Thèse de Doctorat de Mathématiques Appliquées

Dirigée par

Jean-Marc BARDET et William KENGNE

Thèse présentée et soutenue le : 12 Juillet 2021

Après avis des rapporteurs : Christian FRANCQ (ENSAE)
Pascal MASSART (Université Paris-Saclay)

Jury de soutenance :

Jean-Marc BARDET	(Université Paris 1) Directeur de thèse
Fabienne COMTE	(Université de Paris) Examinatrice
Christian FRANCQ	(ENSAE) Rapporteur
William KENGNE	(Cergy Paris Université) Codirecteur
Béatrice Laurent-Bonneau	(INSA) Examinatrice
Pascal MASSART	(Université Paris-Saclay) Rapporteur
Olivier WINTENBERGER	(Sorbonne Université) Examinateur

Résumé

L'analyse des séries temporelles est un sujet de recherche très actif en Statistique et en Data Science. L'abondance de ce type de données a créé d'énormes besoins de méthodologies efficaces et précises. C'est ainsi que plusieurs familles de modèles ont vu le jour. Étant donnée cette multitude de modèles, comment en choisir un pour modéliser une série temporelle ? L'objet de cette thèse est de proposer et d'étudier des critères de sélection de modèles pour une grande famille de modèles contenant les séries temporelles autorégressives telles les ARMA ainsi que les séries temporelles conditionnellement hétéroscédastiques telles les GARCH.

Nous commençons par une brève présentation des séries temporelles, en particulier de la famille des modèles affines causaux tout en rappelant quelques précédents résultats utiles pour cette thèse. Nous décrivons ensuite quelques critères classiques de sélection de modèles obtenus pour les séries temporelles et terminons par un succinct résumé de nos principales contributions.

Dans la suite, nous allons présenter les quatre contributions originales de cette thèse. Le chapitre 2 donne des conditions suffisantes sur la pénalité en fonction de la régularité de la dépendance du processus par rapport à son passé afin d'obtenir un critère consistant en probabilité. Nous proposons également un test d'adéquation du modèle sélectionné basé sur l'autocorrélation du carré des résidus du modèle. Les simulations numériques ont montré des résultats satisfaisants.

Au chapitre 3, nous proposons une généralisation du critère de Hannan et Quinn à la classe des séries affines causales. Cette généralisation induit une certaine constante connue pour les modèles classiques (type ARMA, GARCH ou APARCH) et pour les modèles complexes tels les ARMA-GARCH, la constante est inconnue mais peut être estimée de manière adaptative via l'heuristique de pente. Là également, quelques études de simulation ont attesté de la qualité des critères obtenus.

Dans la troisième contribution, nous construisons des critères asymptotiquement efficients. Nous proposons une généralisation du critère AIC d'Akaike se basant sur la pénalité dite idéale. Le comportement asymptotique de cette pénalité idéale nous a suggéré un terme de pénalité qui vaut exactement $2\,D_m$ comme dans l'AIC pour des modèles assez simples, et pour des modèles complexes, nous avons donné une formule moins explicite. A la suite de Schwartz, nous dérivons également le critère BIC qui s'appuie sur la maximisation de la probabilité a posteriori de choisir le vrai modèle.

Au chapitre 5, nous nous sommes restreints à l'étude non asymptotique d'un processus particulier de la classe des modèles affines causaux. Un estimateur des moindres carrés pénalisé est construit à partir d'un critère de sélection adaptatif et la sélection est opérée parmi une collection de modèles linéaires. Nous avons montré que l'estimateur final est presque aussi performant que le meilleur sur la collection considérée, *i.e.* qu'il réalise, à une constante près, le compromis biais-variance. La pénalité obtenue généralise celle de Mallows et dépend d'une constante que l'on pourrait estimer avec des algorithmes de calibration adaptative.

Enfin, nous donnons quelques pistes de recherche dans la Conclusion générale du travail.

Mots-Clefs : Modèles affines causaux, sélection de modèles, estimation par quasi-vraisemblance, critères consistants, critères efficients, inégalité oracle, heuristique de pente.

Abstract

Time series analysis is a very active research topic in Statistics and Data Science. The abundance of this type of data has created a huge need for efficient and accurate methodologies. Thus, several families of models have emerged. Given this multitude of models, how do you choose one to model a time series? The purpose of this thesis is to propose and study model selection criteria for a large family of models containing autoregressive time series such as ARMA and conditionally heteroscedastic time series such as GARCH.

We start with a brief presentation of time series, in particular of the family of causal affine models, while recalling some previous results useful for this thesis. We then describe some classical model selection criteria obtained for time series and end with a brief summary of our main contributions.

The rest of our work presents our four contributions. The first chapter gives sufficient conditions on the penalty depending on the regularity of the dependence of the process on its past in order to obtain a consistent criterion. We also propose a goodness-of-fit test of the selected model based on the autocorrelation of the square of the model residuals. The numerical simulations have shown satisfactory results.

In Chapter 3, we propose a generalization of the Hannan and Quinn criterion to the class of causal affine series. This generalization induces a certain constant known for classical models (ARMA, GARCH or APARCH type) and can be data-driven estimated for complex models like ARMA-GARCH. Here again, some simulation studies have attested to the quality of the criteria obtained.

In the third contribution, we construct asymptotically efficient criteria. We propose a generalization of Akaike's AIC criterion based on the so-called ideal penalty. The asymptotic behavior of this ideal penalty suggested a penalty term which is exactly $2\,D_m$ as in the AIC for simple models, and for complex models, we give a less explicit formula. Following Schwartz, we also derive the BIC criterion based on the maximization of the a posteriori probability of choosing the true model.

In Chapter 5, we restricted ourselves to the non-asymptotic study of a particular process of the class of causal affines. A penalized least-squares estimator is built on a data driven selected model among a collection of linear models. We showed that the final estimator performs almost as well as the best over the considered collection, *i.e.* it achieves, up to a constant, the bias-variance tradeoff. The penalty obtained generalizes Mallows' penalty and depends on a constant that is estimated with data-driven calibration algorithms.

Finally, we give some research directions in the General Conclusion of the work.

Keywords: Causal affine models, model selection, quasi-likelihood estimation, data-driven, consistent criteria, efficient criteria, oracle inequality, slope heuristic.

Remerciements

La réalisation de ce travail de longue haleine a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais tout d'abord adresser toute ma reconnaissance à mon directeur de thèse, Jean-Marc BARDET, pour sa patience, sa disponibilité, sa rigueur scientifique et surtout ses précieux conseils, qui ont contribué à alimenter ma réflexion. Tu m'as fait confiance dès nos premiers échanges et tu m'as soutenu dans toutes les étapes du concours COFUND Math In Paris qui m'a permis de décrocher une bonne source de financement. JM, 1000 Mercis!

Je tiens à exprimer toute ma gratitude à William KENGNE, co-directeur de cette thèse. Ton dynamisme, ta disponibilité, ta rigueur, tes conseils m'ont été précieux dans l'avancée de mes travaux.

Je suis extrêmement reconnaissant envers Christian FRANCQ et Pascal MASSART pour l'honneur qu'ils m'ont fait en acceptant de rapporter cette thèse. Merci pour cette lecture minutieuse et vos rapports.

Mes remerciements vont à l'endroit de Fabienne COMTE, Béatrice LAURENT-BONNEAU et Olivier WINTENBERGER pour leur participation à mon jury de thèse.

Je remercie la Fondation des Sciences Mathématiques de Paris pour avoir mis sur pied le programme doctoral COFUND Math In Paris. Je dis aussi merci à l'UE pour le financement accordé au COFUND. Je profite pour remercier Ariela pour sa disponibilité et l'accompagnement constant durant les trois années de financement.

J'adresse mes sincères remerciements à Paul DOUKHAN pour les réponses à mes questions.

J'ai une pensée pour mes enseignants du Master MASS POP d'Aix-Marseille, notamment à Marie-Christine ROUBAUD, Nicolas PECH, Thomas WILLER et Pierre PUDLO, pour leur accompagnement durant le concours COFUND. Marie-Christine, c'est aussi grâce à tes paroles motivantes et captivantes qui raisonnent encore aujourd'hui dans mes oreilles que j'ai été lauréat COFUND 2018.

J'exprime également mes remerciements à la chaleureuse équipe du SAMM et plus particulièrement à Nicolas, Branda, Xavier, Dafnis, Florian, Alexandre, Diem, Clara, Alice, Marie, Antoine, Béchir, Alain, Julien, Joseph, la liste étant loin d'être exhaustive.

Une mention spéciale à tous les amis qui n'ont cessé de m'apporter leur soutien: Merci à toi Dolores, Tara (Moussa), NDAM, Marcel, Alida, Ismaël, Ousman.

Je remercie ma famille, ma source de motivation: Maman, chacune de nos discussions me donnait plein d'énergie pour surmonter les moments de doute et travailler sans relâche. Papa, Mami, Digora, Fidel, Bayer merci pour votre soutien.

Notations

$\mathbb{R}^\infty$	Set of sequences of real numbers with a finite number of non-zero terms
$\lvert m\rvert$ or D_m	is the dimension of the model m
p.s.	means almost surely (in Chapter 1)
$X_n \xrightarrow[n\to+\infty]{a.s.} X$	$\mathbb{P}\big(X_n \underset{n\to\infty}{\longrightarrow} X\big) = 1$
$X_n = o_{a.s.}(g_n)$	$\frac{X_n}{g_n} \xrightarrow[n\to+\infty]{a.s.} 0$
$X_n \xrightarrow[n\to\infty]{\mathcal{P}} X$	$\forall \epsilon > 0$, $\mathbb{P}\big(\lvert X_n - X\rvert > \epsilon\big) \underset{n\to\infty}{\longrightarrow} 0$
$X_n = o_P(g_n)$	$\frac{X_n}{g_n} \xrightarrow[n\to\infty]{\mathcal{P}} 0$
$X_n = O_P(g_n)$	$\forall \epsilon > 0$, $\exists M_\epsilon > 0$ such that $\mathbb{P}\big(\lvert X_n\rvert \geq M_\epsilon\, g_n\big) \leq \epsilon$
$X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$	$F_{X_n}(x) \underset{n\to\infty}{\longrightarrow} F_X(x)$ for every number $x \in \mathbb{R}$ at which F is continuous where F_{X_n} and F_X are the cumulative distribution functions of random variables X_n and X, respectively
$X \sim \mathrm{SG}(\sigma^2)$	$\mathbb{E}\big[e^{\lambda X}\big] \leq e^{\lambda^2\sigma^2/2}$ for every $\lambda > 0$ provided that $\mathbb{E}[X] = 0$
$X \sim \mathrm{SE}(\sigma^2, \alpha)$	$\mathbb{E}\big[e^{\lambda X}\big] \leq e^{\lambda^2\sigma^2/2}$, $\forall \lambda : \lvert\lambda\rvert < \frac{1}{\alpha}$ provided that $\mathbb{E}[X] = 0$
$\lVert.\rVert$	denotes the usual Euclidean norm on $\mathbb{R}^\nu$, with $\nu \geq 1$
$\lVert.\rVert$ $\lVert A\rVert = \sup_{v\neq 0} \frac{\lVert Av\rVert}{\lVert v\rVert}$	for a matrix A, denote the subordinate matrix norm such that
$\lVert.\rVert_r$ $\big(\mathbb{E}\big[\lVert X\rVert^r\big]\big)^{1/r} \in [0,\infty]$	if X is a $\mathbb{R}^\nu$-random variable and $r \geq 1$, we set $\lVert X\rVert_r =$
$\lVert.\rVert_\Theta$	for $\theta \in \Theta \subset \mathbb{R}^d$, if $\Psi_\theta : \mathbb{R}^\infty \to E$ where $E = \mathbb{R}^\nu$ or E is a set of square matrix, denote $\lVert\Psi_\theta(\cdot)\rVert_\Theta = \sup_{\theta\in\Theta}\big\{\lVert\Psi_\theta(\cdot)\rVert\big\}$
∂_θ.	for $\theta \in \Theta \subset \mathbb{R}^d$, if $\Psi_\theta : \mathbb{R}^\infty \to \mathbb{R}$ is a $\mathcal{C}^2(\Theta \times \mathbb{R}^\infty)$ function, we will denote $\partial_\theta\Psi_\theta(\cdot) = \Big(\frac{\partial}{\partial\theta_i}\Psi_\theta(\cdot)\Big)_{1\leq i\leq d} = \big(\partial_{\theta_i}\Psi_\theta(\cdot)\big)_{1\leq i\leq d}$
$\partial^2_{\theta^2}$.	$\partial^2_{\theta^2}\Psi_\theta(\cdot) = \Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\Psi_\theta(\cdot)\Big)_{1\leq i,j\leq d}$.

Contents

1

Révue de Littérature et Synthèse des Travaux

Ce chapitre introductif a pour but de présenter tour à tour le modèle statistique étudié, quelques critères de sélection de modèles et nos contributions sur la question de la sélection de modèles pour les séries affines causales.

1.1 Modèles Affines Causaux

1.1.1 Quelques préliminaires sur les séries temporelles

Une série temporelle est une suite d'observations, observées sur une période de temps T. Lorsque la période T est un sous ensemble de $\mathbb{R}$, on parlera de série à temps continu et de série à temps discret quand T est plutôt une partie de $\mathbb{Z}$. Dans toute cette thèse, l'on ne considérera que les séries à temps discret.

Les observations temporelles apparaissent dans de nombreux domaines clés du monde réel, allant de la finance à la biologie, en passant par la météorologie, le traitement du langage naturel, la détection des anomalies (dans les systèmes de contrôle), l'audio et le traitement vidéo, pour n'en citer que quelques-uns. Compte tenu de leur abondance, l'analyse des séries chronologiques est devenue un domaine clé de la Statistique et aussi du Machine Learning.

Dans l'analyse des séries temporelles, l'on s'intèresse à la compréhension et à la modélisation des relations entre les observations d'une même variable: la variable d'intérêt est identique aux variables explicatives à la seule différence qu'elles sont observées à des instants distincts. Donnons une définition plus formelle et générale des séries chronologiques.

Définition 1.1. *Une série temporelle* $(X_1, X_2, \ldots, X_n)$ *est une réalisation du processus stochastique* $(X_t)_{t\in\mathbb{Z}}$ *défini sur un espace probabilisé* $(\Omega, \mathcal{F}, P)$.

À partir d'une seule réalisation $(X_1, X_2, \ldots, X_n)$, l'on aimerait faire de l'inférence sur les paramètres associés au processus $(X_t)_{t\in\mathbb{Z}}$, notamment l'espérance, la variance etc. En Statistique classique, l'on sait que la moyenne empirique des réalisations **indépendantes** d'une même variable aléatoire est un bon estimateur de l'espérance. Et l'on sait aussi que l'hypothèse d'indépendance est fondamentale car si les réalisations sont fortement dépendantes alors il est probable que la moyenne empirique ne recouvre pas l'espace de probabilité tout entier et qu'elle ne converge pas vers l'espérance.
Ainsi, peut-on obtenir de bons estimateurs des paramètres lorsqu'on a affaire à une série temporelle? La réponse est oui mais au prix de quelques conditions sur le processus. Par exemple, l'on peut supposer que le processus est de moyenne constante et qu'il est de mémoire courte de sorte que l'effet de X_t sur X_{t+k} ne soit pas trop important pour k assez grand.
Introduisons plus explicitement la notion de stationnarité qui est un concept assez intuitif et suppose que la structure du processus sous-jacent n'évolue pas avec le temps.

Définition 1.2. *(forte stationnarité) Un processus aléatoire $X = (X_t)_{t\in\mathbb{Z}}$ est dit (fortement) stationnaire s'il est invariant en distribution par toute translation du temps, i.e. $\forall k \in \mathbb{N}^*$, $\forall (t_1, \ldots, t_k) \in \mathbb{Z}^k$, $\forall c \in \mathbb{Z}$, les vecteurs $(X_{t_1}, \ldots, X_{t_k})$ et $(X_{t_1+c}, \ldots, X_{t_k+c})$ ont la même distribution.*

Ce type de stationnarité comme l'indique son nom, est plutôt une condition forte et est souvent difficile à vérifier en pratique. On lui préfère la stationnarité faible qui ne nécessite que l'invariance des moments jusqu'à l'ordre deux.

Définition 1.3. *(faible stationnarité) Un processus aléatoire $X = (X_t)_{t\in\mathbb{Z}}$ est dit (faiblement) stationnaire si son espérance est constante et si pour tout $t, k \in \mathbb{Z}$, la covariance entre X_t et X_{t+k} ne dépend que de k.*

Un autre concept qui accompagne très souvent la notion de stationnarité est l'ergodicité. Il permet de généraliser la loi des grands nombres aux variables aléatoires dépendantes.

Définition 1.4. *(Ergodicité pour processus stationnaire) Le processus fortement stationnaire X est dit ergodique si et seulement si pour tout borélien B et pour tout entier k,*

$$n^{-1} \sum_{t=1}^{n} \mathbb{1}_B(X_t, X_{t+1}, \ldots, X_{t+k}) \xrightarrow[n\to+\infty]{a.s.} \mathbb{P}\big\{(X_1, \ldots, X_{1+k}) \in B\big\}.$$

1.1.2 Modèles affines causaux

Pour décrire les relations entre les observations, l'on recourt à la modélisation. Dans le cadre des séries temporelles, trois grandes familles de modèles se distinguent:

- Les séries linéaires: importantes pour modéliser la structure de covariance de la série. Cette famille comprend les modèles autorégressifs (AR), les modèles de moyenne mobile (MA) et leur combinaison (ARMA);
- Les modèles non linéaires pour prendre en compte les irruptions soudaines et irrationnelles que l'on observe dans les données, notamment financières. Il s'agit des modèles GARCH et associés;
- Les combinaisons des deux premières familles i.e. les ARMA-GARCH.

La classe de modèles affines causaux se propose d'unifier l'écriture de ces trois familles afin de les traiter simultanément dans un cadre identique.

1.1.2.a Modèle général et propriétés

Définition 1.5. *Soit $(\xi_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires iid telle que $\mathbb{E}[|\xi_0|^r] < \infty$. Le processus $(X_t)_{t\in\mathbb{Z}}$ est dit affine causal s'il existe deux fonctions mesurables M, f : $\mathbb{R}^\infty \to \mathbb{R}$ telles que*

$$X_t = M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t + f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \quad \textit{pour tout } t \in \mathbb{Z}. \tag{1.1.1}$$

L'on notera $X = (X_t)_{t\in\mathbb{Z}} \in \mathcal{AC}(M, f)$. Dans toute cette thèse, l'on se restreindra au cas où les fonctions M et f sont connus mais dépendent d'un paramètre inconnu θ^* appartenant à une certaine région admissible de paramètres Θ de $\mathbb{R}^d$. Nos travaux s'inscrivent dans un cadre sémi-paramétrique puisque l'on ne fera aucune hypothèse sur la distribution du bruit $(\xi_t)_{t\in\mathbb{Z}}$.

Une fois le modèle posé, il est fondamental avant l'étape d'inférence, de montrer qu'il est bien défini i.e. qu'il admet de solutions ayant un intérêt pratique (finitude, stationarité, moments, etc). Ainsi, sous quelles conditions (1.1.1) admet-il une solution stationnaire? L'étude de l'existence de solution du modèle (1.1.1) et de certaines propriétés a été effectuée par Doukhan and Wintenberger (2008) en termes de coefficients de contraction des fonctions M_θ et f_θ dits coefficients de Lipschitz.

Nous énonçons la principale condition de Doukhan and Wintenberger (2008) avec $\Psi_\theta = f_\theta, M_\theta$, $H_\theta = M_\theta^2$ où $\theta \in \Theta$ (Θ supposé compact) et $i = 0, 1, 2$.

$\mathbf{A}_i(\Psi_\theta, \Theta)$: Supposons $\|\Psi_\theta(0)\|_\Theta < \infty$ et qu'il existe une suite de nombres positifs $\left(\alpha_k^i(\Psi_\theta, \Theta)\right)_{k\geq 1}$ telle que $\sum_{k=1}^\infty \alpha_k^i(\Psi_\theta, \Theta) < \infty$ vérifiant:

$$\left\| \frac{\partial^i \Psi_\theta(x)}{\partial \theta^i} - \frac{\partial^i \Psi_\theta(y)}{\partial \theta^i} \right\|_\Theta \leq \sum_{k=1}^\infty \alpha_k^i(\Psi_\theta, \Theta)|x_k - y_k| \quad \text{pour tout } x, y \in \mathbb{R}^\infty.$$

Pour les modèles de type GARCH admettant une espérance conditionnelle nulle, il est plus optimale de considérer
$\mathbf{A}_i(H_\theta, \Theta)$: Supposons $\|H_\theta(0)\|_\Theta < \infty$ et qu'il existe une suite de nombres positifs $\left(\alpha_k^i(H_\theta, \Theta)\right)_{k\geq 1}$ telle que $\sum_{k=1}^\infty \alpha_k^i(H_\theta, \Theta) < \infty$ vérifiant:

$$\left\| \frac{\partial^i H_\theta(x)}{\partial \theta^i} - \frac{\partial^i H_\theta(y)}{\partial \theta^i} \right\|_\Theta \leq \sum_{k=1}^\infty \alpha_k^i(H_\theta, \Theta)|x_k^2 - y_k^2| \quad \text{pour tout } x, y \in \mathbb{R}^\infty.$$

L'idée de Doukhan and Wintenberger (2008) consiste à restreindre l'espace des paramètres de sorte qu'une condition sur le moment du bruit entraine l'existence d'une solution stationnaire ergodique admettant des moments. Ainsi, si ξ_0 admet des moments d'ordre $r \geq 1$, considérons

$$\Theta(r) = \Big\{ \theta \in \mathbb{R}^d,\ A_0(f_\theta, \{\theta\}) \text{ et } A_0(M_\theta, \{\theta\}) \text{ vraies avec } \sum_{k=1}^\infty \alpha_k^0(f_\theta, \{\theta\}) + \|\xi_0\|_r \sum_{k=1}^\infty \alpha_k^0(M_\theta, \{\theta\}) < 1 \Big\} \quad (1.1.2)$$

ou bien pour les modèles de type GARCH

$$\Theta(r) = \Big\{ \theta \in \mathbb{R}^d, \quad A_0(H_\theta, \{\theta\}) \text{ vraie avec } \|\xi_0\|_r \sum_{k=1}^\infty \alpha_k^0(H_\theta, \{\theta\}) < 1 \Big\}. \quad (1.1.3)$$

Nous pouvons rappeler le résultat de Doukhan and Wintenberger (2008).

Proposition 1.1. *Pour tout $\theta \in \Theta(r)$ avec $r \geq 1$, (1.1.1) admet une unique solution $(X_t)_{t\in\mathbb{Z}}$ stationnaire, ergodique, faiblement dépendante avec $\|X_0\|_r < \infty$.*

1.1.2.b Exemples de processus affines causaux

1. **Modèle AR(∞)**

   Soit $(\psi_k(\theta))_{k\in\mathbb{N}}$ une suite de réels dépendante de $\theta \in \Theta \subset \mathbb{R}^d$. Considérons l'AR($\infty$) définie ainsi qu'il suit:

   $$X_t = \sum_{k\geq 1} \psi_k(\theta) X_{t-k} + \sigma\, \xi_t \quad \text{pour tout } t \in \mathbb{Z}, \quad (1.1.4)$$

où $(\xi_t)_{t\in\mathbb{Z}}$ est une suite de variable aléatoire iid admettant des moments d'ordre $r \geq 1$, et $\sigma > 0$. Ce processus correspond à (1.1.1) avec $f_\theta\big((x_i)_{i\geq 1}\big) = \sum_{k\geq 1} \psi_k(\theta) x_k$ et $M_\theta \equiv \sigma$ pour tout $\theta \in \Theta$. Les coefficients de Lipschitz associés à f_θ sont $\alpha_k^0(f_\theta) = \|\psi_k(\theta)\|_\Theta$. La Proposition 1.1 est vraie pour tout paramètre θ vérifiant $\sum_{k=1}^\infty \|\psi_k(\theta)\|_{\{\theta\}} < 1$.

Ainsi soit l'ARMA(p, q) suivant

$$X_t - \sum_{i=1}^{p} \phi_i \, X_{t-i} = \xi_t + \sum_{i=1}^{p} \theta_i \, \xi_{t-i}, \tag{1.1.5}$$

tel que toutes les racines du polynôme $\theta(z) = \sum_{i=1}^p \theta_i \, z^i$ soient de module strictement plus grand que 1. Alors le processus $(X_t)_{t\in\mathbb{Z}}$ est inversible et (1.1.5) admet une représentation affine causale (1.1.4).

2. **Modèle ARCH**(∞)

Considérons $(\psi_k(\theta))_{k\in\mathbb{N}}$ une suite de réels dépendante de $\theta \in \Theta \subset \mathbb{R}^d$. Soit l'ARCH$(\infty)$ suivant

$$X_t = \Big(\psi_0(\theta) + \sum_{k=1}^{\infty} \psi_k(\theta) X_{t-k}^2\Big)^{1/2} \xi_t \quad \text{pour tout } t \in \mathbb{Z}, \tag{1.1.6}$$

où $(\xi_t)_t$ admet des moments d'ordre $r \geq 1$. Nous avons là un exemple de (1.1.1) avec $f_\theta = 0$ et $M_\theta\big((x_i)_{i\geq 1}\big) = \Big(\psi_0(\theta) + \sum_{k=1}^\infty \psi_k(\theta) x_{t-k}^2\Big)^{1/2}$. Ainsi les coefficients de Lipschitz (relativement à la série $\{x_t^2\}$) sont $\alpha_k^0(H_\theta) = \|\psi_k(\theta)\|_\Theta$ et tout paramètre θ vérifiant $\|\xi_0\|_r \sum_{k=1}^\infty \|\psi_k(\theta)\|_{\{\theta\}} < 1$ assure que le processus (X_t) est stationnaire, ergodique d'après la Proposition 1.1.

3. **Modèle TARCH**(∞)

Ce modèle est défini ainsi qu'il suit

$$X_t = \Big(\psi_0(\theta) + \sum_{k=1}^{\infty} \psi_k^+(\theta) \max(X_{t-k}, 0) - \psi_k^-(\theta) \min(X_{t-k}, 0)\Big)\xi_t \quad \text{pour tout } t \in \mathbb{Z}, \tag{1.1.7}$$

avec $(\xi_t)_t$ admet des moments d'ordre $r \geq 1$. On a encore un cas de (1.1.1) avec $f_\theta = 0$ et $M_\theta\big((x_i)_{i\geq 1}\big) = \psi_0(\theta) + \sum_{k=1}^\infty \psi_k^+(\theta) \max(X_{t-k}, 0) - \psi_k^-(\theta) \min(X_{t-k}, 0)$. On obtient que $\alpha_k^0(M_\theta) = \max\big(\|\psi_k^+(\theta)\|_\Theta, \|\psi_k^-(\theta)\|_\Theta\big)$ de sorte que tout θ vérifiant $\|\xi_0\|_r \sum_{k=1}^\infty \alpha_k^0(M_\theta) < 1$ assure toutes les propriétés de la Proposition 1.1.

1.1.2.c QMLE et quelques résultats

À présent que nous avons des garanties théoriques sur l'existence de la solution stationnaire du processus (1.1.1), une question naturelle qui vient à l'esprit serait celle de l'estimation du paramètre θ^*.
Supposons que nous avons observé $(X_1, X_2, \ldots, X_n)$ où les X_t proviennent de (1.1.1) avec $\theta^* \in \Theta(r)$ inconnu pour un $r \geq 1$. Nous allons construire un estimateur de θ^* en utilisant un contraste basé sur la vraisemblance.

Dans l'étude des séries chronologiques, très souvent la distribution des innovations est inconnue, c'est pourquoi nous aurons recours à la méthode de quasi-vraisemblance. Elle consiste à supposer dans un premier temps que le bruit est gaussien et à trouver l'expression de la densité jointe des observations. Cette expression qui donne un contraste sera utilisée même si le bruit n'est pas gaussien. Par ailleurs, il est aussi important de rappeler qu'ici, la densité jointe des observations est en pratique impossible à maximiser. L'on se contente d'une version conditionnelle.
Plus formellement, supposons que $(\xi_t)_{t\in\mathbb{Z}}$ soit un bruit blanc gaussien (standard). À partir de (1.1.1), l'on déduit que la log densité de X_t sachant $\sigma\big(X_i, i<t\big)$ est

$$-\frac{1}{2}\Big[\frac{(X_t-f^t_{\theta^*})^2}{H^t_{\theta^*}}+\log(H^t_{\theta^*})\Big].$$

Ainsi, la log densité conditionnelle de $(X_1,\dots,X_n)$ sachant $\sigma\big(X_t, t\leq 0\big)$ est

$$-\frac{1}{2}\sum_{t=1}^{n}\Big[\frac{(X_t-f^t_{\theta^*})^2}{H^t_{\theta^*}}+\log(H^t_{\theta^*})\Big].$$

Desormais, l'hypothèse de gaussianité du bruit est abandonnée et la log-densité conditionnelle permet de définir pour tout $\theta\in\Theta$

$$L_n(\theta):=-\frac{1}{2}\sum_{t=1}^{n}q_t(\theta)\ ,\ \text{avec } q_t(\theta):=\frac{(X_t-f^t_\theta)^2}{H^t_\theta}+\log(H^t_\theta) \tag{1.1.8}$$

où $f^t_\theta:=f_\theta(X_{t-1},X_{t-2},\cdots)$, $M^t_\theta:=M_\theta(X_{t-1},X_{t-2},\cdots)$ et $H^t_\theta=\big(M^t_\theta\big)^2$. L'on remarque tout de suite que la fonction de vraisemblance L_n n'est pas évaluable puisqu'elle dépend du passé $(X_{-j})_{j\in\mathbb{N}}$ qui est inconnu. C'est pourquoi, nous considèrerons une approximation observable de L_n notée $\widehat{L}_n$ et définie ainsi qu'il suit

$$\widehat{L}_n(\theta):=-\frac{1}{2}\sum_{t=1}^{n}\widehat{q}_t(\theta)\ ,\ \text{avec } \widehat{q}_t(\theta):=\frac{(X_t-\widehat{f}^t_\theta)^2}{\widehat{H}^t_\theta}+\log(\widehat{H}^t_\theta) \tag{1.1.9}$$

où $\widehat{f}^t_\theta:=f_\theta(X_{t-1},X_{t-2},\dots,X_1,0,\dots,0)$, $\widehat{M}^t_\theta:=M_\theta(X_{t-1},X_{t-2},\dots,X_1,0,\dots,0)$ et $\widehat{H}^t_\theta=\big(\widehat{M}^t_\theta\big)^2$. L'estimateur de quasi-vraisemblance (QMLE) de θ^* est donc:

$$\widehat{\theta}_n=\underset{\theta\in\Theta}{\operatorname{argmax}}\ \widehat{L}_n(\theta). \tag{1.1.10}$$

Notons que les propriétés (consistance et normalité asymptotique) de cet estimateur ont été étudiées par Bardet and Wintenberger (2009) sous certaines hypothèses de régularité qui seront utilisées tout au long de cette thèse.

La première hypothèse porte sur l'identifiabilité du "vrai" paramètre θ^*. Il induit que deux paramètres qui conduisent à une même valeur de la vraisemblance sont identiques.

A1: *Pour tout* $\theta,\theta'\in\Theta$, $(f^0_\theta=f^0_{\theta'}$ *et* $M^0_\theta=M^0_{\theta'})$ *p.s.* $\implies\theta=\theta'$.

Notons que cette hypothèse est vérifiée pour tous les modèles affines causaux classiques lorsque le bruit est non dégénéré.

Aussi, dans les définitions des fonctions de vraisemblance et quasi-vraisemblance conditionnelles, il apparait un dénominateur qui ne devrait pas s'annuler. Ainsi, nous supposerons

A2: $\exists \underline{h} > 0$ *tel que* $\inf_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ *pour tout* $x \in \mathbb{R}^\infty$.

La condition suivante permet d'assurer l'inversibilité de la matrice hessienne de la fonction L_n et est très importante pour prouver la normalité asymptotique de l'estimateur QMLE.

A3: *Une des familles* $(\partial f_\theta^t/\partial\theta^{(i)})_{1\leq i\leq d}$ *ou* $(\partial H_\theta^t/\partial\theta^{(i)})_{1\leq i\leq d}$ *est linéairement indépendante presque sûrement.*

Théorème 1.1. *(Bardet and Wintenberger (2009)) Sous les conditions* ***A1-A3*** *et si*

$$\alpha_k^0(f_\theta,\Theta) + \alpha_k^0(M_\theta,\Theta) = O(k^{-\gamma}) \quad avec\, \gamma > 3/2,$$

Alors le QMLE est fortement consistent i.e.

$$\widehat{\theta}_n \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} \theta^*. \tag{1.1.11}$$

Si par ailleurs,

$$\alpha_k^1(f_\theta,\Theta) + \alpha_k^1(M_\theta,\Theta) = O(k^{-\gamma}) \quad avec\, \gamma > 3/2,$$

Alors $\widehat{\theta}_n$ *est asymptotiquement normal i.e.*

$$\sqrt{n}\,(\widehat{\theta}_n - \theta^*) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}\Big(0, \big(F(\theta^*)\big)^{-1} G(\theta^*) \big(F(\theta^*)\big)^{-1}\Big) \tag{1.1.12}$$

avec $G(\theta^*) = \frac{1}{4}\,\mathbb{E}\Big[\partial_\theta q_0(\theta^*)\,\partial_\theta q_0(\theta^*)^\top\Big]$ *et* $F(\theta^*) = -\frac{1}{2}\,\mathbb{E}\big[\partial^2_{\theta^2} q_0(\theta^*)\big]$.

1.2 Sélection de modèles

1.2.1 Pourquoi sélectionner un modèle?

En pratique, pour une analyse, le statisticien dispose des données. À partir de ces données, il aimerait tirer des informations pour atteindre l'objectif de l'étude. L'extraction de ces informations passe par une phase de modélisation. Quel est le modèle le plus approprié pour ces données? Rappelons qu'une mauvaise modélisation conduira à des conclusions gravement trompeuses et donc à une prise de décision catastrophique puisque l'intérêt pratique de la Statistique est d'éclairer les décideurs. Comme il n'existe pas de modèle qui soit universellement adapté à toutes les données, le statisticien explore plusieurs modèles/algorithmes afin de mieux comprendre le processus ayant généré ces données ou d'obtenir une meilleure performance de prédiction.
Ainsi, il est beaucoup plus judicieux de considérer une grande famille de modèles pour réduire le risque de se tromper et de choisir le modèle le plus approprié de la collection.

La sélection de modèles est donc le processus de sélection d'un modèle dans une famille convenablement choisie à cet effet étant donné un jeu de données. Elle est fondamentale à toute analyse de données pour des fins d'inférence ou de prédiction fiable. Elle intervient donc dans tous les domaines utilisant les données notamment l'économie, l'ingénierie, l'écologie, la finance, les sciences politiques, la biologie, l'épidémiologie, etc.

1.2.2 Objectifs de la Sélection de modèles

L'objectif d'une procédure de sélection de modèles est en réalité le but de la modélisation. On en distingue principalement deux:

- L'identification ou l'interprétation: ici, l'on aimerait comprendre le processus de génération des données, interpréter la nature des données ou appuyer un modèle physique. L'on préfèrera donc les modèles assez simples et interprétables aux modèles complexes;
- La prédiction: Il ne s'agit plus de découvrir ou d'inférer des modèles beaucoup plus réels mais plutôt de répondre à des questions pratiques, comme: quelles températures il fera demain, quel sera le cours du bitcoin dans deux semaines, comment détecter une fraude lors d'une transaction, etc.

1.2.2.a Identification

Considérons $\mathcal{M}$, la famille de modèles candidats. Puisque l'objectif est de sélectionner le vrai modèle (notée m^* dans tout ce travail) ayant généré les données, l'on fait l'hypothèse que $\mathcal{M}$ contient m^*.

Une procédure de sélection de modèles vise à sélectionner le meilleur modèle de la famille noté $\widehat{m} := \widehat{m}_n$ dans le but d'estimer m^*.

Définition 1.6. *Une telle procédure sera dite* ***consistante*** *si elle arrive à retrouver m^* avec probabilité approchant 1 asymptotiquement. Plus formellement,*

$$\mathbb{P}\big(\widehat{m} = m^*\big) \underset{n\to\infty}{\longrightarrow} 1. \tag{1.2.1}$$

Il existe aussi une version de la consistance avec une convergence presque sûre. Elle est dite **consistance forte**:

$$\widehat{m} \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} m^*. \tag{1.2.2}$$

La première est beaucoup plus classique dans la littérature et est facilement vérifiable à travers des études de Monte Carlo.

1.2.2.b Prédiction

Bien que la consistance d'une procédure de sélection de modèles est une propriété mathématique convaincante, en pratique l'on ne connait pas le vrai modèle et supposer que $m^* \in \mathcal{M}$ est souvent irréaliste. Le vrai modèle peut être de dimension infinie et il est impossible d'identifier pareil modèle. Dans ce cas, ce qui pourrait être fait est une approximation de la réalité par des modèles de dimension finie qui sont évidemment tous faux. Et le but serait donc de sélectionner le modèle qui fait le moins possible d'erreur quand il s'agit de prédire. Ainsi, l'on va considérer une collection de modèles au plus dénombrable $\mathcal{M}_n$ de cardinal dépendant de n. Pour un certain prédicteur θ, l'on mesure son erreur de prédiction par

$$R(\theta) := \mathbb{P}\gamma(\theta) = \mathbb{E}[\gamma(\theta, X_1)]$$
$$\text{avec} \quad \gamma(\theta, X_t) := \frac{(X_t - f_\theta^t)^2}{H_\theta^t} + \log(H_\theta^t) \tag{1.2.3}$$

Le contraste γ est -2 fois la log-densité conditionnelle de X_t sachant le passé (en ayant considéré un bruit gaussien). D'après Bardet and Wintenberger (2009), la fonction R

atteint son minimum en θ^* qui est souvent non estimable directement soit parcequ'il est de dimension infinie ou parce que la famille de modèle candidat est mal spécifiée. Ainsi, il est naturel de définir un risque de prédiction relativement à θ^*

$$\ell(\theta, \theta^*) := R(\theta) - R(\theta^*) \geq 0. \tag{1.2.4}$$

Le meilleur choix possible dans la collection $\mathcal{M}_n$ est appelé *oracle* (que nous noterons aussi m^* mais qui ne signifie plus vrai modèle) défini par

$$m^* \in \arg \inf_{m \in \mathcal{M}_n} \ell(\widehat{\theta}_m, \theta^*). \tag{1.2.5}$$

L'oracle m^* est un idéal et n'est pas atteignable puisqu'il dépend de θ^* et de la distribution du processus qui sont inconnus. Ainsi, une procédure de sélection de modèles peut avoir pour but de sélectionner un modèle $\widehat{m}$ qui imitera m^* en matière de risque.

Définition 1.7. *Une procédure de sélection de modèles est dite asymptotiquement* ***efficiente*** *si le rapport de risque de $\widehat{m}$ et de m^* converge en probabilité vers 1 quand $n \longrightarrow \infty$; i.e.,*

$$\frac{\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)}{\ell(\widehat{\theta}_{m^*}, \theta^*)} \xrightarrow[n\to\infty]{\mathcal{P}} 1. \tag{1.2.6}$$

L'estimateur $\widehat{\theta}_m$ n'est plus le QMLE mais le minimiseur du risque empirique associée au contraste γ:

$$\widehat{\theta}_m = \underset{\theta \in \Theta_m}{\operatorname{argmin}}\, \gamma_n(\theta) \quad \text{avec} \quad \gamma_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \gamma(\theta, X_t)$$

Notons que la version non asymptotique de cette définition est la propriété de plus en plus recherchée chez les théoriciens de la sélection de modèles: elle est appelée **inégalité oracle**

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1 \inf_{m \in \mathcal{M}_n} \left\{ \ell(\widehat{\theta}_m, \theta^*) \right\} + \frac{C_2}{n} \tag{1.2.7}$$

avec C_1 proche de 1 et $C_2 > 0$. Le but est d'obtenir (1.2.7) sur un évènement de grande probabilité ou en espérance.

1.2.3 Stratégies de Sélection de modèles

Nous avons vu l'intérêt d'un recours à la sélection de modèles, les différents objectifs. Mais comment est opérée cette sélection? Cette sous-section se propose de répondre à cette question.

1.2.3.a Tests Statistiques

Le problème de sélection de modèles peut être traité en utilisant un test d'hypothèses. Dans cette approche, l'on compare les modèles candidats deux à deux si ceux-ci sont imbriqués, sinon l'on pourrait préalablement les plonger dans un modèle plus grand les contenant. Ainsi, un test de Fisher (F-test) peut être réalisé pour départager les deux modèles. Pour faire simple, supposons que nous voulons sélectionner entre un AR(1) et un AR(2) ayant observé une trajectoire. Comme hypothèses du test, l'on peut considérer pour hypothèse nulle "$\theta^* \in \Theta_1 \subset]-1, 1[$ " vs "$\theta^* \in \Theta_2 \subset]-1, 1[\times]-1, 1[$" l'hypothèse alternative. Un rejet de l'hypothèse nulle pourrait conduire à retenir l'AR(2) bien que cela

ne signifie pas que H1 est vraie. Ainsi, cela pourrait servir pour progréssivement éliminer certains modèles.

Pour conserver un sous-ensemble de paramètres importants, l'on dispose également des tests comme le test de Wald, de student ou du rapport de vraisemblance.
Cette approche de sélection de modèles par test présente un certain nombre de défauts: avant d'effectuer le test, il faut se fixer un niveau de significativité; ainsi, deux analystes différents avec exactement les mêmes modèles et données peuvent arriver à des conclusions différentes basées sur des niveaux de significativité différents. Aussi, le nombre de F-test a effectué croit très vite.

1.2.3.b Validation croisée

La validation croisée est très utilisée notamment en apprentissage statistique/ Machine Learning quand il s'agit d'un objectif de prédiction. Plusieurs approches de la validation croisée existent (Stone (1974), Allen (1974)). L'on distingue principalement deux :

- La **testset validation** ou **hold out**: Elle consite à diviser les données en deux: la première partie appélée données d'entrainement servira à calculer les estimations des paramètres des modèles et la deuxième appelée données test permettra d'évaluer les capacités prédictives des estimateurs obtenus lors de la phase d'apprentissage. Le choix de la proportion de la partie qui servira de test (généralement moins de 40%) influence beaucoup le résultat final;
- La k **fold validation** qui consiste à diviser les données en k blocs et d'utiliser $k-1$ blocs pour entrainer les estimateurs et le bloc restant pour le test. Cet entrainement est repété k fois de sorte que chaque bloc serve exactement une fois d'ensemble test. Contrairement au hold-out, le choix de k est certes important mais n'affectera pas sensiblement le résultat final.

En pratique, il est connu que le hold-out est relativement instable par rapport au k fold. Une question assez ouverte est de savoir comment choisir k et quelle est son influence sur les performances statistiques de la méthode? A cet effet, Arlot (2007) a montré qu'il y'a surpénalisation lorsque k est petit et d'un point de vue asymptotique, il est indispensable de faire tendre k vers l'infini avec n pour obtenir une procédure optimale.

Plusieurs autres études ont abordé cette approche de sélection de modèles dans le cadre de l'estimation de la densité (voir par exemple Celisse et al. (2014), Celisse (2008)).

1.2.3.c Pénalisation

Une approche assez populaire pour opérer une sélection de modèles est la pénalisation d'un certain critère empirique qui peut être -2 fois la log-vraisemblance ou le contraste des moindres carrés. Puisque le critère empirique est généralement décroissant en fonction de la complexité du modèle, le minimiser fournira le modèle le plus complexe qui sera difficilement interprétable et en plus n'aura pas nécessairement les meilleures capacités de prédiction en raison de sa grande variabilité due à l'estimation d'un très grand nombre de paramètres.

L'idée de pénaliser vient donc résoudre ce problème; en ajoutant un terme (croissant en fonction de la complexité), l'on défavorise les très gros modèles. A l'inverse, si ce terme est prépondérant devant le contraste empirique, l'on va privilégier les modèles beaucoup plus simples présentant un biais beaucoup plus important.

La pénalisation remonte aux années 1970 avec les travaux de Akaike (1969), Mallows (1973) et Akaike (1973). Bien qu'il soit probable que cette idée ait déjà existée dans

d'autres contextes tels que la sélection de sous-ensembles par Beale et al. (1967), Hocking and Leslie (1967), et la régression ridge par Hoerl and Kennard (1970).
En utilisant les moindres carrés ordinaires dans le cadre de régression, Mallows a obtenu le critère C_p. Parallèlement, Akaike a dérivé l'AIC pour l'estimation de la densité en utilisant le contraste log-vraisemblance. Quelques années plus tard, à la suite d'Akaike, Schwarz (1978) a proposé une approche alternative pour l'estimation de la densité et a dérivé le critère d'information bayésien (BIC).

Le terme de pénalité de ces critères est proportionnel à la dimension du modèle. Au cours des dernières décennies, différentes approches de pénalisation ont émergé telles que la pénalisation $\mathbb{L}^2$ pour la régularisation ridge Hoerl and Kennard (1970), la pénalisation $\mathbb{L}^1$ pour la procédure LASSO de Tibshirani (1996) et l'elastic net qui combine les normes $\mathbb{L}^1$ et $\mathbb{L}^2$ Zou and Hastie (2005).

Formellement, en considérant comme contraste empirique la quasi-vraisemblance, l'on définit un critère de pénalisation $\widehat{C}$ ainsi qu'il suit:

$$\widehat{C}(m) = -2\,\widehat{L}_n\big(\widehat{\theta}(m)\big) + \kappa_n(m), \tag{1.2.8}$$

pour tout modèle candidat $m \in \mathcal{M}$, avec κ_n est une suite croissante de la taille de modèle. Puisqu'il existe une multitude de choix pour κ_n, toute la difficulté de la sélection de modèles par pénalisation réside dans la spécification de ce terme de pénalité. Les choix les plus classiques sont entre autres:

- $\kappa_n(m) = 2c\,D_m \log\log n$ avec $c > 1$, on retrouve le critère de Hannan et Quinn Hannan and Quinn (1979);
- $\kappa_n = D_m \log n$, $\widehat{C}$ est le critère BIC Schwarz (1978);
- $\kappa_n = 2\,D_m$, $\widehat{C}$ est l'AIC Akaike (1973).

Toutes ces pénalités ont été obtenues soit pour la prédiction ou l'interprétation. Rappelons aussi qu'il existe aussi des critères de pénalisation où le terme de pénalité est considéré multiplicativement par rapport au risque empirique; c'est l'exemple du FPE de Akaike.

Compte tenu de cette richesse de choix, comment un statisticien peut-il décider du critère à utiliser? C'est la raison pour laquelle les approches qui ne peuvent être ni mises en œuvre ni comprises par la communauté scientifique ne sont pas acceptées. Cela implique qu'il faut au moins une méthode qui puisse être implémentée facilement et donnent des résultats qui peuvent être interprétés par les utilisateurs. D'un point de vue statistique, les approches suffisamment générales pour traiter une grande variété de problèmes et vérifiant une propriété de sélection de modèles sont les plus prisées.

1.2.4 Dérivation de quelques critères dans le cadre des séries temporelles

1.2.4.a FPE Akaike (1969)

C'est l'un des tout premiers critères de pénalisation à être obtenus du moins pour les séries chronologiques (notamment pour un AR(p)). Supposons avoir observé une trajectoire $(X_1, \ldots, X_n)$ d'un AR(p) $X_t = \sum_{i=1}^{p} \phi_i^* X_{t-i} + \sigma\,\xi_t$. Puisque l'on ignore la valeur de p, alors quel ordre d'auto-régression choisir pour ajuster les données? L'idée d'Akaike est de choisir l'ordre qui minimise l'erreur quadratique moyenne lorque le modèle estimé (au moyen de $(X_1, \ldots, X_n)$) est utilisé pour prédire la valeur Y_{n+1} d'une suite d'observation $(Y_1, \ldots, Y_n)$ provenant du même AR(p) mais indépendant de $(X_1, \ldots, X_n)$.

Soit $\widehat{\phi} := \big(\widehat{\phi}_1(X_1,\ldots,X_n),\ldots,\widehat{\phi}_p(X_1,\ldots,X_n)\big)^\top$, l'estimateur du QMLE de $\phi^* = (\phi_1^*,\ldots,\phi_p^*)^\top$. D'après le TCL (1.1.12), on a

$$\sqrt{n}\,(\widehat{\phi}-\phi^*) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\big(0,(G(\phi^*))^{-1}\big) \tag{1.2.9}$$

puisque $G(\phi^*) = -F(\phi^*)$. Pour $i,j = 1,\ldots,p$

$$\begin{aligned}(G(\phi^*))_{i,j} &= \frac{1}{4}\,\mathbb{E}\Big[\frac{\partial q_0(\phi^*)}{\partial\phi_i}\frac{\partial q_0(\phi^*)}{\partial\phi_j}\Big]\\ &= \frac{1}{4}\,\mathbb{E}\Big[\frac{4}{\sigma^4}\,X_{-i}X_{-j}\xi_0^2\Big] = \frac{1}{\sigma^2}\,\mathbb{E}\big[X_i\,X_j\big].\end{aligned}$$

Avec $X = (X_1,\ldots,X_n)^\top$ et $Y = (Y_1,\ldots,Y_n)^\top$, l'erreur quadratique est donc:

$$\begin{aligned}\mathbb{E}\big[(Y_{n+1}-\widehat{\phi}_1\,Y_n-\cdots-\widehat{\phi}_p\,Y_{n+1-p})^2\big] &= \sigma^2\,\mathbb{E}[\xi_{n+1}^2] + \mathbb{E}\left[\Big(\sum_{i=1}^{p}(\widehat{\phi}_i-\phi_i^*)\,Y_{n+1-i}\Big)^2\right]\\ &= \sigma^2 + \mathbb{E}\Big[(\widehat{\phi}-\phi^*)^\top\,\mathbb{E}\big[Y\,Y^\top|X\big]\,(\widehat{\phi}-\phi^*)\Big]\\ &= \sigma^2 + \sigma^2\,\mathbb{E}\Big[(\widehat{\phi}-\phi^*)^\top\,G(\phi^*)\,(\widehat{\phi}-\phi^*)\Big]\\ &\approx \sigma^2 + \sigma^2\,\frac{1}{n}\,\mathrm{Trace}\Big(G(\phi^*)\,(G(\phi^*))^{-1}\Big)\\ &= \sigma^2\Big(1+\frac{p}{n}\Big)\end{aligned} \tag{1.2.10}$$

où l'approximation a été rendue possible grâce à (1.2.9). Puisque σ est en général inconnu, l'idée est de le substituer par un estimateur consistant. D'après Brockwell and Davis (1991), l'estimateur MLE de σ^2 vérifie

$$n\,\frac{\widehat{\sigma}^2}{\sigma^2} \xrightarrow[n\to\infty]{\mathcal{D}} \chi(n-p)$$

avec $\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\Big(X_t - \sum_{i=1}^{p}\widehat{\phi}_i\,X_{t-i}\Big)^2$. En remplaçant σ^2 par l'estimateur $\frac{n}{n-p}\widehat{\sigma}^2$ dans la forme dévéloppée (1.2.10), Akaike définit le FPE pour un modèle candidat d'ordre p quelconque à être

$$\mathrm{FPE}_p = \widehat{\sigma}^2\frac{n+p}{n-p}.$$

Ing and Wei (2005) ont démontré que le FPE possède la propriété d'efficience asymptotique.

1.2.4.b AIC Akaike (1973)

L'AIC a été obtenu par Akaike (1973). Il fut dérivé en recherchant un estimateur non biaisé de la divergence de Kullback entre le vrai modèle et les modèles candidats. Supposons observé $X_1,\ldots,X_n$ d'un ARMA(p,q)

$$X_t - \sum_{i=1}^{p}\phi_i^*\,X_{t-i} = \xi_t + \sum_{i=1}^{q}\beta_i^*\,\xi_{t-i}$$

où (ξ_t) est un bruit blanc gaussien. Posons $\theta^* = (\phi^*,\beta^*)$, la divergence de Kullback-Leibler entre la loi de X_0 (sachant tout le passé) suivant le vrai modèle (représenté par θ^*) et un modèle candidat est

$$\mathcal{K}(\theta^*|\theta) = \mathbb{E}\Big[-2\,\ln\Big(\frac{g(X_0,\theta)}{g(X_0,\theta^*)}\Big)\Big]$$

où l'espérance est prise sous θ^* et $g(X_0, \theta)$ représente la densité de X_0 sachant $X_{-1}, X_{-2}, \ldots$ sous θ. L'objectif est de construire un estimateur non biaisé de $\mathcal{K}(\theta^*|\theta)$. Puisque cette pseudo distance est positive et s'annule uniquement en $\theta = \theta^*$, minimiser cet estimateur non biaisé nous conduira au modèle le "plus proche" du vrai.

Pour simplifier, supposons que les modèles compétitifs sont aussi gaussiens. Puisque $\mathbb{E}\big[-2\,\ln\big(g(X_0, \theta^*)\big)\big]$ est indépendant du modèle candidat choisi, nous ignorerons ce terme dans l'expression de $\mathcal{K}(\theta^*|\theta)$. Donc, un estimateur naturel de cette approximation sera :

$$-\frac{2}{n}\sum_{t=1}^{n}\ln\big(g(X_t,\theta)\big) = -\frac{2}{n}\,L_n(\theta).$$

Considérons $\widehat{\theta}$ l'estimateur MLE de θ^*. Sous l'hypothèse de différentiabilité de L_n et en prenant $\widehat{\theta}$ comme maximum local, un développement de Taylor au second ordre nous donne

$$L_n(\theta^*) \approx L_n(\widehat{\theta}) + \frac{1}{2}\,(\theta^* - \widehat{\theta})^\top\,\partial^2_{\theta^2} L_n(\widehat{\theta})\,(\theta^* - \widehat{\theta}). \tag{1.2.11}$$

En vertu du théorème ergodique et de la continuité uniforme (voir Bardet and Wintenberger (2009)), l'on a:

$$\frac{1}{n}\,\partial^2_{\theta^2} L_n(\widehat{\theta}) \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} F(\theta^*) = -\frac{1}{2}\,\mathbb{E}\big[\partial^2_{\theta^2} q_0(\theta^*)\big]$$

avec $q_t(\theta) = -2\,\ln\big(g(X_t,\theta)\big)$. Le TCL (1.1.12) s'écrit ici

$$\sqrt{n}\,(\widehat{\theta} - \theta^*) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}\Big(0, 2\,\big(\mathbb{E}\big[\nabla^2_\theta q_0(\theta^*)\big]\big)^{-1}\Big) \tag{1.2.12}$$

puisque $G(\theta^*) = \frac{1}{4}\,\mathbb{E}\Big[\partial_\theta q_0(\theta^*)\,\partial_\theta q_0(\theta^*)^\top\Big] = -F(\theta^*)$.

Ainsi, l'on obtient en prenant l'espérance de chaque terme en (1.2.11)

$$\begin{aligned}\mathbb{E}\big[L_n(\widehat{\theta})\big] &\approx \mathbb{E}\big[L_n(\theta^*)\big] - \frac{1}{2}\,\mathrm{Trace}\Big(F(\theta^*)\,2\,\big(\mathbb{E}\big[\nabla^2_\theta q_0(\theta^*)\big]\big)^{-1}\Big)\\ &= \mathbb{E}\big[L_n(\theta^*)\big] + \frac{1}{2}\,(p+q).\end{aligned} \tag{1.2.13}$$

En effectuant également un développement de $\mathbb{E}\big[q_0(\theta)\big]$ autour de θ^*, il vient que

$$\mathbb{E}\big[q_0(\widehat{\theta})\big] \approx \mathbb{E}\big[q_0(\theta^*)\big] + \frac{1}{2}\,(\widehat{\theta} - \theta^*)^\top\,\mathbb{E}\big[\nabla^2_\theta q_0(\theta^*)\big]\,(\widehat{\theta} - \theta^*),$$

le terme de premier ordre étant nul puisque θ^* est l'unique minimum de la divergence de Kullback. Ainsi,

$$\mathbb{E}\Big[\mathbb{E}\big[q_0(\widehat{\theta})\big]\Big] \approx \mathbb{E}\big[q_0(\theta^*)\big] + \frac{p+q}{n}. \tag{1.2.14}$$

En prenant l'espérance de la différence entre les termes de (1.2.13) et (1.2.14), on a

$$\begin{aligned}\mathbb{E}\Big[-\frac{2}{n}\,L_n(\widehat{\theta}) - \mathbb{E}\big[q_0(\widehat{\theta})\big]\Big] &\approx \mathbb{E}\big[-\frac{2}{n}\,L_n(\theta^*)\big] - \mathbb{E}\big[q_0(\theta^*)\big] - \frac{p+q}{n} - \frac{p+q}{n}\\ &= -2\frac{(p+q)}{n}\end{aligned}$$

de sorte qu'un estimateur sans biais (non normalisé) de $\mathbb{E}\big[q_0(\widehat{\theta})\big]$ est

$$AIC = -2\,L_n(\widehat{\theta}) + 2\,(p+q).$$

L'optimalité asymptotique de l'AIC a été prouvée par Shibata (1980) et aussi par Ing and Wei (2005).

1.2.4.c BIC Schwarz (1978)

Le critère BIC est l'un des plus connus en sélection de modèles. Contrairement aux autres critères, sa dérivation s'est faite dans un cadre bayésien: le paramètre θ^* associé au vrai modèle m^* est supposé aléatoire. Etant donné une famille de modèle $\mathcal{M}$, un modèle $m \in \mathcal{M}$ est tiré selon une distribution a priori π_m. Très souvent cet a priori est non informatif car pris comme distribution uniforme ou bien ignoré dans la formule d'approximation du BIC. Ainsi, conditionnellement au modèle m tiré, un paramètre $\theta \in \Theta_m$ est tiré suivant une autre distribution μ_m. Le choix du meilleur modèle dans la famille $\mathcal{M}$ est opéré en considérant le modèle maximisant la probabilité a posteriori i.e.,

$$\begin{aligned}\widehat{m} &= \underset{m\in\mathcal{M}}{\operatorname{argmax}} \left\{\mathbb{P}\big(m \,|\, X_1, X_2, \ldots, X_n\big)\right\} \\ &= \underset{m\in\mathcal{M}}{\operatorname{argmax}} \left\{\frac{\pi_m\, \mathbb{P}\big(X_1, X_2, \ldots, X_n \,|m\big)}{\mathbb{P}\big(X_1, X_2, \ldots, X_n\big)}\right\}.\end{aligned}$$

Le terme prépondérant dans cette formule est bien $\mathbb{P}\big(X_1, X_2, \ldots, X_n \,|m\big)$ qui peut encore s'écrire en intégrant sur l'espace des paramètres

$$\begin{aligned}\mathbb{P}\big(X_1, X_2, \ldots, X_n \,|m\big) &= \int_{\Theta_m} \mathbb{P}\big(X_1, X_2, \ldots, X_n \,|\theta, m\big)\, d\mu_m(\theta) \\ &= \int_{\Theta_m} \exp\big(L_n(\theta)\big)\, d\mu_m(\theta).\end{aligned}$$

Cette intégrale est généralement impossible à calculer et est approchée au moyen de la formule de l'approximation de Laplace (Lebarbier and Mary-Huard (2004)). Ainsi, l'on obtient

$$\log \mathbb{P}\big(X_1, X_2, \ldots, X_n \,|m\big) \approx L_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m| + \frac{\log(2\pi)}{2}\,|m| - \frac{1}{2}\,\log\big(\det\big(-\partial^2_\theta L_n(\widehat{\theta})\big)\big) + O(n^{-1}).$$

Et en prenant $\pi_m = 1/|\mathcal{M}|$, il vient que

$$\begin{aligned}\log\Big(\pi_m\, \mathbb{P}\big(X_1, X_2, \ldots, X_n \,|m\big)\Big) \approx L_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m| \\ \underbrace{-\log(|\mathcal{M}|) + \frac{\log(2\pi)}{2}\,|m| - \frac{1}{2}\,\log\big(\det\big(-\partial^2_\theta L_n(\widehat{\theta})\big)\big) + O(n^{-1})}_{O_P(1)}.\end{aligned}$$

En négligeant tous ces termes qui restent bornés (et le terme $\mathbb{P}\big(X_1, X_2, \ldots, X_n\big)$ indépendant du choix de m) , il vient que maximiser la probabilité a posteriori revient à maximiser $L_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m|$. Ce qui conduit Schwartz à définir le BIC à être

$$BIC_m = -2\, L_n(\widehat{\theta}_m) + \log(n)|m|$$

et donc à choisir le modèle $\widehat{m}$ comme celui qui minimise le BIC.

Plusieurs études ont montré que si le vrai modèle m^* est inclus dans la famille des modèles candidats $\mathcal{M}$, alors $\widehat{m}$ obtenu par minimisation du BIC est consistant (voir par exemple Csiszár and Shields (1999), Csiszár et al. (2000), Garivier (2006), Bardet et al. (2012)).

1.2.4.d HQ Hannan and Quinn (1979)

Il a été obtenu par Hannan and Quinn (1979) dans le cadre de l'estimation de l'ordre à considérer lorsqu'un modèle linéaire autorégressif est ajusté aux données. Le terme de pénalité en $\log\log n$ provient d'une application de la loi des logarithmes itérérés (LIL) à la série des autocorrélations partielles. Il est défini par

$$HQ = -2\,L_n(\widehat{\theta}) + 2\,c\,p\,\log\log n$$

où $c > 1$. En effet, Hannan and Quinn (1979) ont argué que le $\log n$ dans la pénalité BIC qui assure la consistance dans la sélection du vrai ordre, n'est pas la suite qui croit la plus lentement possible et ont proposé le $\log\log n$ qui vérifie la propriété de forte consistance Hannan and Quinn (1979), Hannan and Deistler (2012). D'après eux, le BIC peut très souvent choisir des modèles très simples, éventuellement erronés, pour de petits échantillons.

1.2.5 Calibration de la constante multiplicative

Dans la littérature de la sélection de modèles non asymptotique, il est fréquent de rencontrer les pénalités dépendantes d'une certaine constante universelle inconnue:

$$\mathrm{pen}(m) = \kappa\,\mathrm{pen}_{\mathrm{shape}}(m)$$

où $\mathrm{pen}_{\mathrm{shape}}$ est une fonction de la dimension du modèle

- D_m pour les modèles linéaires gaussiens Birgé and Massart (2001) ;
- $D_m\,(1+\sqrt{2\,L_m})$ Birgé and Massart (2007b) dans un cadre gaussien plus général;
- $D_m\left(1 + c\,\log(n/D_m)\right)$ pour la détection de ruptures Lebarbier (2005);
- et bien d'autres.

Déterminer la valeur de la constante κ est fondamentale avant la mise en œuvre de la procédure de sélection. Massart (2007) a proposé une méthode efficace pour calibrer κ au moyen des données uniquement. Il s'agit de l'*heuristique de pente* qui se décline en deux algorithmes classiques:

- *slope estimation*: elle consiste à prendre comme une estimation de κ deux fois la valeur de la pente de croissance du critère empirique en fonction de la dimension D_m pour les modèles complexes (voir Figure 1.1);
- *saut de dimension* qui consite à considérer plutôt 2 fois la valeur de κ (en ayant préalablement choisi une grille de valeurs pour κ) qui donne le plus grand saut de la compléxité (voir Figure 1.2).

1.3 Synthèse des travaux

1.3.1 Critères de Sélection de modèles consistants et test d'ajustement pour les modèles affines causaux

Supposons avoir observé une trajectoire $(X_1, X_2, \ldots, X_n)$ du processus affine causal

$$X_t = M_{\theta^*}\left((X_{t-i})_{i\in\mathbb{N}^*}\right)\xi_t + f_{\theta^*}\left((X_{t-i})_{i\in\mathbb{N}^*}\right) \text{ pour tout } t\in\mathbb{Z} \tag{1.3.1}$$

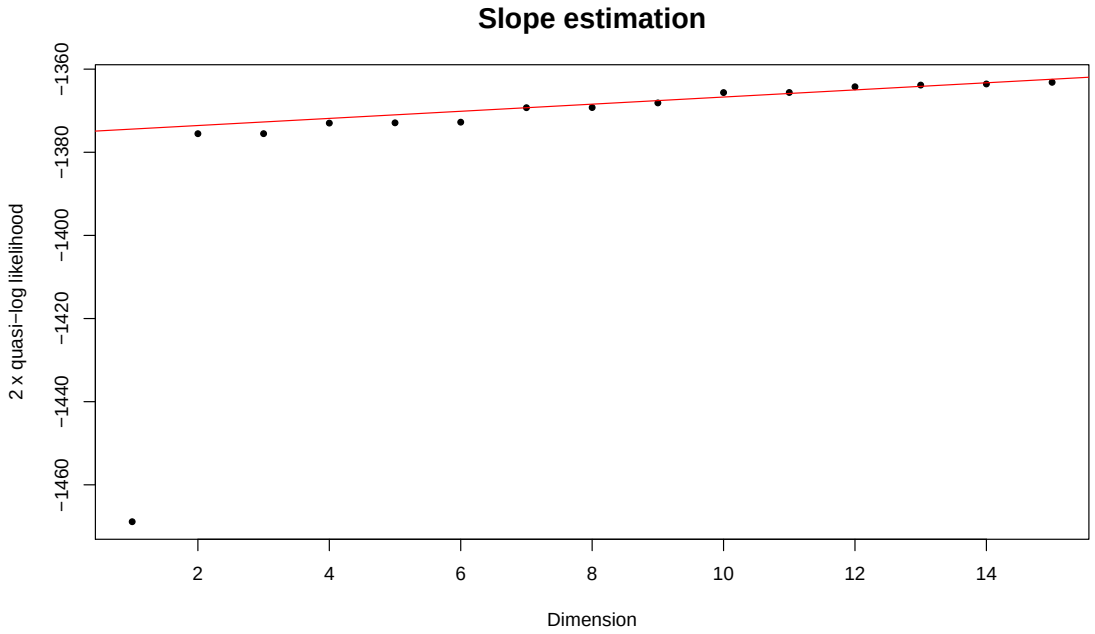


Figure 1.1: Courbe de la quasi log-vraisemblance d'un AR(2) en fonction de la dimension; la pente de la droite oblique vaut $\widehat{a} = 0.86$.

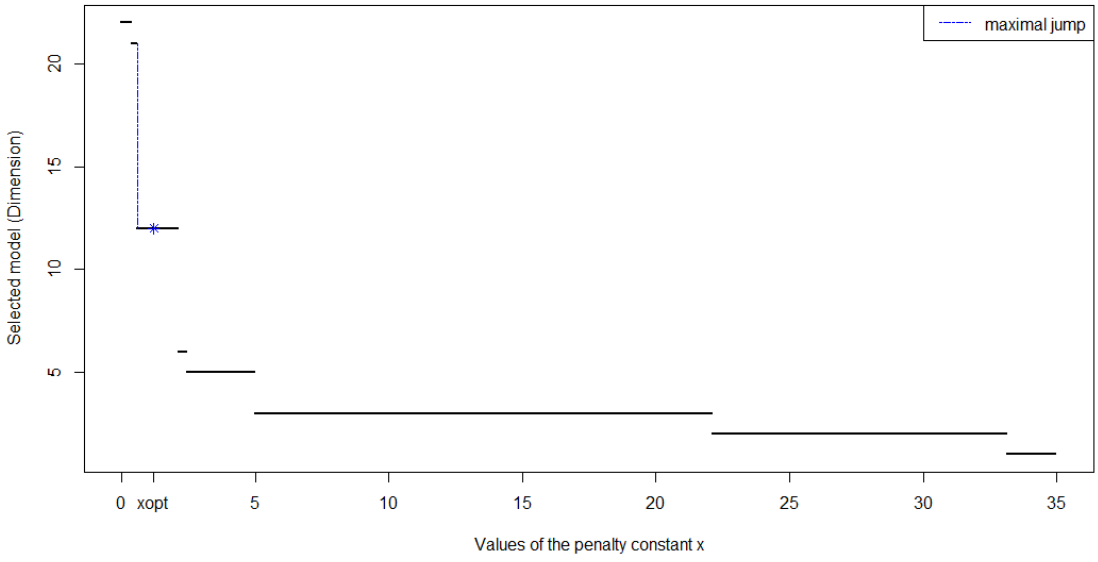


Figure 1.2: Dimension Jump

où $\theta^* \in \mathbb{R}^d$. Le modèle (1.3.1) ayant généré les données sera noté m^* et appartient à une certaine famille finie de modèles affines causaux $\mathcal{M}$. Notre objectif est de construire des critères de sélection vérifiant la propriété de consistance (1.2.1) et de tester la qualité d'ajustement du modèle sélectionné aux données. Pour ce faire, nous construisons un estimateur $\widehat{m}$ de m^* par pénalisation de la quasi-vraisemblance:

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\, \widehat{C}(m) \quad \text{avec} \quad \widehat{C}(m) = -2\,\widehat{L}_n\big(\widehat{\theta}(m)\big) + |m|\,\kappa_n, \tag{1.3.2}$$

A la suite de Bardet et al. (2012), l'idée est de trouver des conditions sur la suite (κ_n) afin de garantir la convergence (1.2.1); lesquelles conditions devraient dépendre des coefficients de Liptschitz des fonctions M_θ et f_θ.

Avant d'énoncer les principaux résultats obtenus, nous donnons un résultat intermédiaire pratique qui dit qu'il est toujours possible de plonger deux modèles affines causaux paramétriques dans un plus grand modèle affine causal.

Proposition 1.2. *Soit $d_1, d_2 \in \mathbb{N}$, $\Theta_1 \subset \mathbb{R}^{d_1}$ et $\Theta_2 \subset \mathbb{R}^{d_2}$, et pour $i = 1, 2$, définissons $f^{(i)}_{\theta_i}, M^{(i)}_{\theta_i} : \mathbb{R}^\infty \to \mathbb{R}$ et pour $\theta_i \in \Theta_i$. Alors, il existe $\max(d_1, d_2) \le d \le d_1 + d_2$, $\Theta \subset \mathbb{R}^d$, et une famille de fonctions $f_\theta : \mathbb{R}^\infty \to \mathbb{R}$ et $M_\theta : \mathbb{R}^\infty \to [0, \infty)$ avec $\theta \in \Theta$, tels que pour tout $\theta_1 \in \Theta_1$ et $\theta_2 \in \Theta_2$, il existe $\theta \in \Theta$ vérifiant*

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

Ainsi dans toute la suite, Θ sera l'espace des paramètres contenant tous les Θ_m avec $m \in \mathcal{M}$.

1.3.1.a Hypothèses

Outre les hypothèses de régularité ayant permis d'obtenir le TCL (1.1.12), nous considérerons la condition suivante nous garantissant une relation entre la suite κ_n et la vitesse de décroissance des coefficients de Lipschitz de M_θ et f_θ.

Hypothèse $\boldsymbol{K}(\Theta)$: *Sous les hypothèses $A_i(\Psi_\theta, \Theta)$, avec $i = 0, 1$, $\Psi_\theta = f_\theta, M_\theta$ et s'il existe $r \ge 2$ tel que $\theta^* \in \Theta(r)$. Par ailleurs, avec $s = \min(1, r/3)$, supposons que la suite $(\kappa_n)_{n\in\mathbb{N}}$ satisfasse*

$$\sum_{k\ge 1} \Big(\frac{1}{\kappa_k}\Big)^s \Big(\sum_{j\ge k} \alpha^0_j(f_\theta, \Theta) + \alpha^0_j(M_\theta, \Theta) + \alpha^1_j(f_\theta, \Theta) + \alpha^1_j(M_\theta, \Theta)\Big)^s < \infty.$$

Pour les modèles de type GARCH, nous considérerons plutôt

Hypothèse $\widetilde{\boldsymbol{K}}(\Theta)$: *Sous les hypothèses $A_i(\Psi_\theta, \Theta)$, avec $i = 0, 1$, $\Psi_\theta = f_\theta, M_\theta$ et supposons qu'il existe $r \ge 2$ tel que $\theta^* \in \Theta(r)$. Par ailleurs, avec $s = \min(1, r/4)$, supposons que la suite $(\kappa_n)_{n\in\mathbb{N}}$ satisfasse*

$$\sum_{k\ge 1} \Big(\frac{1}{\kappa_k}\Big)^s \Big(\sum_{j\ge k} \alpha^0_j(\widetilde{H}_\theta, \Theta) + \alpha^1_j(\widetilde{H}_\theta, \Theta)\Big)^s < \infty.$$

Remarque 1.1. *Ces conditions sur $(\kappa_n)_{n\in\mathbb{N}}$ ne sont pas restrictives : par exemple, si les coefficients de Lipschitz de f_θ, M_θ (le cas utilisant $\widetilde{H}_\theta$ peut être traité de manière similaire) et leurs dérivées sont bornées par une décroissance géométrique ou riemanienne, l'on a:*

1. *Cas géometrique:* $\alpha_j^0(f_\theta,\Theta)+\alpha_j^0(M_\theta,\Theta)+\alpha_j^1(f_\theta,\Theta)+\alpha_j^1(M_\theta,\Theta)=O(a^j)$ *avec* $0\le a<1$, *alors tout* (κ_n) *telle* $1/\kappa_n=o(1)$ *peut être choisie; par exemple* $\kappa_n=\log n$ *ou* $\log(\log n)$*; C'est le cas pour les ARMA, GARCH, APARCH or ARMA-GARCH.*

2. *Cas Riemanien:* $\alpha_j^0(f_\theta,\Theta)+\alpha_j^0(M_\theta,\Theta)+\alpha_j^1(f_\theta,\Theta)+\alpha_j^1(M_\theta,\Theta)=O(j^{-\gamma})$ *avec* $\gamma>1$*:*

   - *si* $r\ge 3$ *alors*
     - *si* $\gamma>2$ *alors toute suite* (κ_n) *vérifiant* $1/\kappa_n=o(1)$ *peut être choisie;*
     - *si* $1<\gamma<2$*, toute suite vérifiant* (κ_n) *telle que* $\kappa_n=O(n^\delta)$ *avec* $\delta>2-\gamma$ *peut être choisie.*
   - *si* $1\le r<3$
     - *si* $\gamma>(r+3)/r$ *alors toute suite* (κ_n) *satisfaisant* $1/\kappa_n=o(1)$ *peut être choisie;*
     - *si* $1<\gamma<(r+3)/r$ *alors toute suite* (κ_n) *vérifiant* $\kappa_n=n^\delta$ *avec* $\delta>(r+3)/r-\gamma$ *peut être choisie.*

     Dans le dernier cas de ces deux conditions sur r*, nous pouvons voir que le choix habituel du BIC,* $\kappa_n=\log n$ *ne vérifie pas l'hypothèse en général.*

1.3.1.b Résultats théoriques

Théorème 1.2. *Considérons* $(X_1,\dots,X_n)$ *une trajectoire observée d'un processus affine causal* X *appartenant à* $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$ *(ou* $\widetilde{\mathcal{AC}}(\widetilde{H}_{\theta^*})$*) où* θ^* *est un paramètre inconnu du compact* Θ *contenu dans* $\Theta(r)\subset\mathbb{R}^d$ *(ou* $\widetilde{\Theta}(r)\subset\mathbb{R}^d$*) avec* $r\ge 4$*. Si les hypothèses* ***A1-A3*** *et* $K(\Theta)$ *(ou* $\widetilde{K}(\Theta)$*),* $A_2(f_\theta,\Theta)$ *et* $A_2(M_\theta,\Theta)$ *(ou* $A_2(\widetilde{H}_\theta,\Theta)$*) sont satisfaites, alors*

$$\mathbb{P}(\widehat{m}=m^*)\underset{n\to\infty}{\longrightarrow}1\quad et\quad \widehat{\theta}(\widehat{m})\underset{n\to\infty}{\xrightarrow{\mathcal{P}}}\theta^*. \tag{1.3.3}$$

Nous avons aussi établi la normalité asymptotique du QMLE du modèle choisi.

Théorème 1.3. *Sous les hypothèses du Théorème 1.2 et si* $\theta^*\in\overset{\circ}{\Theta}$*, alors*

$$\sqrt{n}\left(\big(\widehat{\theta}(\widehat{m})\big)_i-(\theta^*)_i\right)_{i\in m^*}\underset{n\to\infty}{\xrightarrow{\mathcal{D}}}\mathcal{N}_{|m^*|}\big(0\ ,\ F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}\big) \tag{1.3.4}$$

où $\big(F(\theta^*,m^*)\big)_{i,j}=-\frac{1}{2}\mathbb{E}\Big[\frac{\partial^2 q_0(\theta^*)}{\partial\theta_i\partial\theta_j}\Big]$ *et* $(G(\theta^*,m^*))_{i,j}=\frac{1}{4}\mathbb{E}\Big[\frac{\partial q_0(\theta^*)}{\partial\theta_i}\frac{\partial q_0(\theta^*)}{\partial\theta_j}\Big]$ *pour* $i,j\in m^*$*.*

1.3.1.c Test Portmanteau

Bien que les pénalités obtenues nous garantissent la convergence des critères, il est également important de vérifier si le modèle choisi est approprié. Cette section tente de répondre à cette question en construisant un test portmanteau comme outil de diagnostic basé sur la séquence des carrés des résidus du modèle choisi. Pour tout $m\in\mathcal{M}$, et K un entier naturel non nul, soit le vecteur corrélogramme des carrés des résidus suivant

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

où pour $k=1,\dots,K$, $\widehat{\rho}_k(m):=\dfrac{\widehat{\gamma}_k(m)}{\widehat{\gamma}_0(m)}$ avec

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

Le résultat suivant fournit un TCL pour $\widehat{\rho}(m^*)$ et $\widehat{\rho}(\widehat{m})$ et établit la distribution asymptotique de la statistique de test.

Théorème 1.4. *Sous les hypothèses du Théorème 1.3, avec en outre*

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

alors,

1. *avec $V(\theta^*,m^*)$ définie en* (1.3.7)*, on a*
$$\sqrt{n}\,\widehat{\rho}(m^*) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\big(0\,,\,V(\theta^*,m^*)\big). \tag{1.3.5}$$
2. *avec $\widehat{Q}_K(m^*) := n\,\widehat{\rho}(m^*)'\big(V(\widehat{\theta}(m^*),m^*)\big)^{-1}\widehat{\rho}(m^*)$, nous avons*
$$\widehat{Q}_K(m^*) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(K). \tag{1.3.6}$$
3. *Les points 1. et 2. demeurent vraies lorque m^* est remplacé par $\widehat{m}$.*

$$V(\theta^*,m^*) := I_K + (\mu_4-1)^{-2}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J_K'(m^*) - 2\,(\mu_4-1)^{-1}\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\,J_K'(m^*). \tag{1.3.7}$$

Le Théorème 1.4 permet de tester asymptotiquement:

$$\begin{cases} H_0:\ \exists m^*\in\mathcal{M}, \text{ tel que } (X_1,\dots,X_n) \text{ est une trajectoire de } X\in\mathcal{AC}(M_{\theta^*},f_{\theta^*}) \\ H_1:\ \nexists m^*\in\mathcal{M}, \text{ tel que } (X_1,\dots,X_n) \text{ est une trajectoire de } X\in\mathcal{AC}(M_{\theta^*},f_{\theta^*}) \end{cases}.$$

avec $\theta^*\in\Theta(m^*)$ dans les deux cas.

1.3.2 Généralisation du critère de Hannan et Quinn pour les séries affines causales

Nous considérons toujours le problème de sélection de modèles (1.3.2). Nous avons vu précédemment qu'il est important de faire dépendre la suite (κ_n) des coefficients de Lipschitz de M_θ et f_θ. Une limite à cela est d'intérêt pratique: puisque nous donnons un large évantail de choix de la pénalité, comment l'analyste choisit-il sa pénalité? En décroissance exponentielle par exemple, parmi $\log(\log n)$, $\log n$ ou même $\sqrt{n}$, laquelle est la plus optimale?

Dans cette nouvelle contribution, nous établissons pour les processus à mémoire pas trop longue typiquement

$$\alpha_j^0(f_\theta,\Theta)+\alpha_j^0(M_\theta,\Theta)+\alpha_j^1(f_\theta,\Theta)+\alpha_j^1(M_\theta,\Theta)=O(j^{-\gamma}) \quad \text{avec} \quad \gamma>2, \tag{1.3.8}$$

que l'on peut toujours trouver une constante c telle que avec $\kappa_n = 2\,c\,\log(\log n)$, l'on ait la propriété de forte consistance (1.2.2). La constante c est connue pour les modèles classiques ARMA, GARCH, APARCH mais pour les modèles complexes i.e. combinaison des ARMA et GARCH, il est difficile de trouver la valeur théorique de c. Cependant, les algorithmes de calibration de la constante multiplicative peuvent être utilisés pour une estimation adaptative de c dans les cas complexes.

Ainsi, la condition $\boldsymbol{K}(\Theta)$ devient

Condition $\boldsymbol{K}(\Theta)$:

$$\sum_{k\geq e}\frac{1}{\log\log k}\sum_{j\geq k}\alpha_j^0(f_\theta,\Theta)+\alpha_j^0(M_\theta,\Theta)+\alpha_j^1(f_\theta,\Theta)+\alpha_j^1(M_\theta,\Theta)<\infty.$$

Et pour les modèles de type GARCH, l'on considèrera

Condition $\widetilde{\boldsymbol{K}}(\Theta)$:

$$\sum_{k\geq e}\frac{1}{\log\log k}\sum_{j\geq k}\alpha_j^0(H_\theta,\Theta)+\alpha_j^1(H_\theta,\Theta)<\infty.$$

Il est facile de voir que tout processus satisfaisant (1.3.8) vérifie $\boldsymbol{K}(\Theta)$.

Pour obtenir un résultat assez général (avec les ARMA-GARCH inclus), il nous a été commode de supposer une relation entre la matrice d'information de Fisher $G(\theta_m^*)$ et la matrice limite de la matrice hessienne de la log-likelihood $F(\theta_m^*)$

$$\big(F(\theta_m^*)\big)_{i,j}=-\frac{1}{2}\,\mathbb{E}\Big[\frac{\partial^2 q_0(\theta_m^*)}{\partial\theta_i\partial\theta_j}\Big]\quad\text{et}\quad(G(\theta_m^*))_{i,j}=\frac{1}{4}\,\mathbb{E}\Big[\frac{\partial q_0(\theta_m^*)}{\partial\theta_i}\frac{\partial q_0(\theta_m^*)}{\partial\theta_j}\Big],$$

avec $\theta_m^*:=(\theta^*,0,\ldots,0)^\top\in\Theta(m)$.

A4: Il existe des constantes α_1 et α_2 telles que pour tout $m\in\mathcal{M}$ vérifiant $m^*\subset m$,

$$\mathbf{1}_m^\top\Sigma_{\theta_m^*}\mathbf{1}_m=\alpha_1\,D_m^1+\alpha_2\,D_m^2 \tag{1.3.9}$$

où D_m^1 et D_m^2 sont tels que $D_m^1+D_m^2=D_m$, $\quad\mathbf{1}_m:=(1,1,\ldots,1)^\top\in\mathbb{R}^{D_m}$ et $\Sigma_{\theta_m^*}:=G(\theta_m^*)^{1/2}F(\theta_m^*)^{-1}G(\theta_m^*)^{1/2}$.
Nous verrons que **A4** est vérifiée pour tous les modèles qui ne sont pas combinaison de ARMA et GARCH.

1.3.2.a Résultats théoriques

La proposition suivante suggère l'existence d'un terme qui sera capital pour notre résultat principal.

Proposition 1.3. *Considérons $(X_1,X_2,\ldots,X_n)$ une trajectoire observée d'un processus affine causal m^*. Pour tout modèle m vérifiant $\theta_m^*\in\widetilde{\overset{\circ}{\Theta}(m)}$, et si **A1-A4** sont vérifiées, alors on a*

$$\limsup_{n\to\infty}\frac{\widehat{L}_n\big(\widehat{\theta}(m)\big)-\widehat{L}_n(\theta_m^*)}{2\,\log\log n}=\frac{1}{4}\big(\alpha_1\,D_m^1+\alpha_2\,D_m^2\big)\quad p.s. \tag{1.3.10}$$

Soit $m\in\mathcal{M}$, désignons par $c_{min}(m)$ la quantité

$$c_{min}(m):=\frac{1}{4}\big(\alpha_1\,D_m^1+\alpha_2\,D_m^2\big) \tag{1.3.11}$$

Nous énonçons maintenant un résultat qui donne les valeurs de α_1 et de α_2 pour la plupart des modèles causaux affines classiques.

Proposition 1.4. *Sous les hypothèses et notation de la Proposition 1.3, nous avons*

- *Si* $\mu_4 = \mathbb{E}[\xi_0^4] = 3$ *(bruit gaussien par exemple), alors* $\alpha_1 = 2$, $\alpha_2 = 2$ *et* $c_{min}(m) = \frac{1}{2} D_m$*;*
- *Si le paramètre* θ *identifiant un modèle affine causal* $X_t = M_\theta^t \xi_t + f_\theta^t$ *peut s'écrire* $\theta = (\theta_1, \theta_2)'$ *avec* $f_\theta^t = \widetilde{f}_{\theta_1}^t$ *et* $M_\theta^t = \widetilde{M}_{\theta_2}^t$*, alors* $\alpha_1 = 2$, $\alpha_2 = \mu_4 - 1$ *et*

$$c_{min}(m) = \frac{1}{2} D_m^1 + \frac{\mu_4 - 1}{4} D_m^2$$

La seconde configuration de la Proposition 1.4 contient la plupart des séries temporelles classiques

- modèles GARCH(p, q), APARCH(δ, p, q) et associés, $c_{min}(m) = \frac{\mu_4 - 1}{4} D_m$;
- modèles ARMA(p, q), $c_{min}(m) = \frac{D_m}{2}$ si la variance du bruit est connue et $c_{min}(m) = \frac{D_m - 1}{2} + \frac{\mu_4 - 1}{4}$ sinon.

Le principal résultat de notre contribution peut donc être énoncé.

Théorème 1.5. *Considérons* $(X_1, \ldots, X_n)$ *une trajectoire observée d'un processus affine causal* X *appartenant à* $\mathcal{AC}(M_{\theta^*}, f_{\theta^*})$ *(ou* $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$*) où* θ^* *est un paramètre inconnu du compact* Θ *contenu dans* $\Theta(r) \subset \mathbb{R}^d$ *(ou* $\widetilde{\Theta}(r) \subset \mathbb{R}^d$*) avec* $r \geq 8$*. Supposons les hypothèses* ***A1-A3****,* $K(\Theta)$ *(ou* $\widetilde{K}(\Theta)$*),* $A_2(f_\theta, \Theta)$ *et* $A_2(M_\theta, \Theta)$ *(ou* $A_2(\widetilde{H}_\theta, \Theta)$*) satisfaites. Soit* $\mathcal{M}$ *une famille finie de modèles contenant* m^* *et vérifiant* ***A4****. Alors, avec* $c_{min} := \max\left(\frac{\alpha_1}{2}, \frac{\alpha_2}{2}\right)$*, on a que*

pour tout $\kappa_n(m) = 2\, c\, D_m \log\log n$ *avec*

$$c \geq c_{min}, \tag{1.3.12}$$

le modèle sélectionné $\widehat{m}$ *par* (1.3.2) *vérifie*

$$\widehat{m} \xrightarrow[n\to\infty]{a.s.} m^*. \tag{1.3.13}$$

Ce résultat nous donne une généralisation de la pénalité fortement consistance de Hannan et Quinn obtenue pour les processus linéaires.

1.3.3 Sélection de modèles efficiente et consistante pour les séries temporelles

Les deux premières contributions se sont focalisées sur la propriété de consistance pour la classe des séries affines causales. À présent, nous allons nous intéresser aux propriétés d'efficience notamment asymptotique.

Soit $(X_1, X_2, \ldots, X_n)$ une trajectoire observée du processus (1.3.1). L'objectif n'est plus de retrouver le vrai modèle m^*.

Nous avons défini en (1.2.3) le risque de prédiction par θ de l'observation X_t au moyen d'un contraste basé sur la log-densité de X_t sachant le passé. Ainsi, le contraste empirique associé est

$$\gamma_n(\theta) := \mathbb{P}_n \gamma(\theta, .) = \frac{1}{n} \sum_{t=1}^{n} \gamma(\theta, X_t),$$

avec $\gamma(\theta, X_t) = q_t(\theta)$ et

$$\mathbb{P}_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{X_t},$$

où δ_{X_t} est la distribution de Dirac de l'observation X_t. Pour les mêmes raisons évoquées sur la nécessité d'approximer la fonction L_n, l'on définit

$$\widehat{\gamma}_n(\theta) = \mathbb{P}_n\widehat{\gamma}(\theta, .) = \frac{1}{n}\sum_{t=1}^{n} \widehat{\gamma}(\theta, X_t),$$

où $\widehat{\gamma}(\theta, X_t) = \widehat{q}_t(\theta)$.

Considérons $\mathcal{M}$ une famille finie de modèles. L'on aimerait construire des critères de sélection de sorte à obtenir l'efficience asymptotique (1.2.6). Pour cela, pour tout $m \in \mathcal{M}$, l'on définit le minimiseur du risque empirique

$$\widehat{\theta}_m = \underset{\theta\in\Theta_m}{\operatorname{argmin}}\, \widehat{\gamma}_n(\theta). \tag{1.3.14}$$

Définissons la pénalité pen : $m \in \mathcal{M} \mapsto \text{pen}(m) \in \mathbb{R}^+$. Considérons

$$\widehat{m}_{\text{pen}} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{\widehat{C}_{\text{pen}}(m)\right\} \qquad \text{avec} \qquad \widehat{C}_{\text{pen}}(m) := \widehat{\gamma}_n\big(\widehat{\theta}_m\big) + \text{pen}(m). \tag{1.3.15}$$

Pour atteindre l'oracle (1.2.5), la pénalité idéale à considérer en (1.3.15) est

$$\text{pen}_{id}(m) = R(\widehat{\theta}_m) - \widehat{\gamma}_n(\widehat{\theta}_m). \tag{1.3.16}$$

Cependant, cette pénalité comme son nom l'indique n'est pas du tout accessible. Dans un premier temps, nous chercherons à étudier sa distribution asymptotique afin d'y extraire une pénalité "proche" de celle idéale. Pour cela, il est indispensable d'établir un TCL général qui sera également valable pour des modèles ne contenant pas m^*. Outre les hypothèses **A1-A3** considérées précédemment, nous supposerons:

1.3.3.a Hypothèses

A4: *Pour tout $m \in \mathcal{M}$, si $(\overline{\theta}_{m,n})$ est une suite de Θ_m vérifiant $\overline{\theta}_{m,n} \xrightarrow[n\to+\infty]{a.s.} \theta^*_m$, alors*

$$\limsup_{n\to\infty} \left\{\mathbb{E}\Big[\Big(\Big\|\frac{1}{n}\left(\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_m)\right)_{i,j\in m}\Big)^{-1}\Big\|^8\Big]\right\} < \infty. \tag{1.3.17}$$

Remarque 1.2. *Notons que dès que les fonctions $\theta \to M_\theta$ et $\theta \to f_\theta$ sont $\mathcal{C}^2(\Theta)$ et si $\overline{\theta}_{m,n} \xrightarrow[n\to+\infty]{a.s.} \theta^*_m$ alors*

$$\Big\|\Big(\frac{1}{n}\left(\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_{m,n})\right)_{i,j\in m}\Big)^{-1}\Big\|^8 \xrightarrow[n\to+\infty]{a.s.} \Big\|\Big(\big(-\frac{1}{2}\,\partial^2_{\theta_i\theta_j}\gamma(\theta^*_m)\big)_{i,j\in m}\Big)^{-1}\Big\|^8.$$

Ainsi, à partir du théorème d'Egorov, on peut trouver un événement $\widetilde{\Omega}$ avec une probabilité suffisamment grande pour que la relation (1.3.17) dans l'hypothèse **A4** *tienne si l'on prend l'espérance sur l'événement $\widetilde{\Omega}$. Pour le cas particulier des processus linéaires, l'hypothèse* **A4** *est vraie sous une condition légère sur la distribution de X, voir par exemple Papangelou (1994) et Findley and Wei (2002).*

Nous supposerons enfin que la vitesse de décroissance de $(\alpha_j(f_\theta,\Theta))_j$, $(\alpha_j(M_\theta,\Theta))_j$, $(\alpha_j(\partial_\theta f_\theta,\Theta))_j$ et $(\alpha_j(\partial_\theta M_\theta,\Theta))_j$ doivent être suffisamment rapides pour garantir la forte consistance et la normalité asymptotique de $\widehat{\theta}_m$:

A5: *Supposons* $A_i(\Psi_\theta,\Theta)$, $\Psi_\theta = f_\theta, M_\theta$ *avec*

$$\alpha_j^0(f_\theta,\Theta)+\alpha_j^0(M_\theta,\Theta)+\alpha_j^1(f_\theta,\Theta)+\alpha_j^1(M_\theta,\Theta)=O(j^{-\delta}) \quad où \quad \delta>7/2.$$

Notons que l'hypothèse A5 ne permet pas de considérer des processus à mémoire longue, mais les séries temporelles causales habituelles à mémoire courte satisfont à cette hypothèse.

1.3.3.b Résultats théoriques

Le résultat suivant montre que le TCL (1.1.12) s'étend même si le modèle candidat m est mal spécifié.

Théorème 1.6. *Soit* $(X_1,\dots,X_n)$ *une trajectoire observée d'un processus affine causal* X *appartenant à* $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$ *(ou* $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$*) où* θ^* *est un paramètre inconnu du compact* Θ *contenu dans* $\Theta(r)\subset\mathbb{R}^d$ *(ou* $\widetilde{\Theta}(r)\subset\mathbb{R}^d$*) avec* $r\geq 8$*. Supposons aussi* **A1**-**A5** *satisfaites, alors pour tout* $m\in\mathcal{M}$*,*

$$\sqrt{n}\left((\widehat{\theta}_m)_i-(\theta_m^*)_i\right)_{i\in m}\underset{n\to\infty}{\xrightarrow{\mathcal{D}}}\mathcal{N}\Big(0\,,\,\big(F(\theta_m^*)\big)^{-1}G(\theta_m^*)\big(F(\theta_m^*)\big)^{-1}\Big), \tag{1.3.18}$$

Ainsi, nous pouvons asymptotiquement approximer (1.3.16).

Proposition 1.5. *Sous les hypothèses du Théorème 1.6, il existe une suite bornée* $(v_n^*)_{n\in\mathbb{N}^*}$ *telle que pour tout* $m\in\mathcal{M}$*, l'on ait*

$$\mathbb{E}\big[pen_{id}(m)\big]\underset{n\to\infty}{\sim}-\frac{2}{n}\,Trace\Big(\big(F(\theta_m^*)\big)^{-1}G(\theta_m^*)\Big)+\frac{v_n^*}{n}. \tag{1.3.19}$$

L'on verra par la suite que le terme en trace dans (1.3.19) est connu pour la plupart des modèles classiques.
Nous énonçons un résultat de consistance qui montre que si la pénalité tends en probabilité vers 0, alors le critère $\widehat{C}_{\rm pen}$ ne sélectionne pas asymptotiquement un modèle mal spécifié.

Théorème 1.7. *Sous les hypothèses du Théorème 1.6, supposons aussi qu'il existe* $\varepsilon>0$ *tel que,*

$$n\,\mathbb{P}\big(\mathrm{pen}(m)\geq\varepsilon\big)\underset{n\to\infty}{\longrightarrow}0 \quad pour\ tout\ m\in\mathcal{M}. \tag{1.3.20}$$

alors,

$$n\,\mathbb{P}\big(m^*\not\subset\widehat{m}_{\rm pen}\big)\underset{n\to\infty}{\longrightarrow}0. \tag{1.3.21}$$

À présent, nous pouvons spécifier la vitesse de convergence de pen à considérer afin d'obtenir un excès de risque proche de celui de l'oracle.

Théorème 1.8. *Sous les hypothèses du Théorème 1.6 et si pour tout* $\varepsilon>0$*, il existe* $K_\varepsilon>0$ *tel que*

$$\mathbb{P}\Big(\limsup_{n\to\infty}\max_{m\in\mathcal{M}}\big(n\,\mathrm{pen}(m)\big)\geq K_\varepsilon\Big)\leq\varepsilon. \tag{1.3.22}$$

Alors, pour tout $\varepsilon>0$*, il existe* $M_\varepsilon>0$ *et* $N_\varepsilon\in\mathbb{N}^*$ *tels que pour tout* $n\geq N_\varepsilon$*,*

$$\mathbb{P}\Big(\ell(\widehat{\theta}_{\widehat{m}_{\rm pen}},\theta^*)\leq\inf_{m\in\mathcal{M}}\big\{\ell(\widehat{\theta}_m,\theta^*)\big\}+\frac{M_\varepsilon}{n}\Big)\geq 1-\varepsilon. \tag{1.3.23}$$

Remarque 1.3. *Remarquons que cette optimalité asymptotique est assez différente de l'optimalité classique de l'efficience asymptotique, où le cardinal de la collection $\mathcal{M}$ et la dimension des modèles compétitifs dépendent de n. Cependant, cette dernière est généralement étudiée dans le cadre où le paramètre θ^* est de dimension infinie (voir par exemple Shibata (1980), Li (1987), Hsu et al. (2019b))*

1.3.3.c Dérivation du BIC pour les modèles affines causaux

Le résultat suivant nous donne une approximation de $\widehat{S}_n(m, X) := \log\big(\pi_m \, \mathbb{P}\big(X \,|\, m\big)\big)$ qui est le terme principal dans la probabilité a posteriori de choisir le modèle m. En gardant les notations de la sous-section 1.2.4, nous supposons qu'il existe une fonction borélienne positive $\theta \to b_m(\theta)$ telle que $d\mu_m(\theta) = b_m(\theta)\, d\theta$.

Théorème 1.9. *Sous les hypothèses* **A1, A2, A3, A5** *et si pour tout $x \in \mathbb{R}^\infty$, les fonctions $\theta \to M_\theta$ et $\theta \to f_\theta$ sont $\mathcal{C}^6(\Theta)$ et satisfont $A_k(f_\theta, \Theta)$ et $A_k(M_\theta, \Theta)$ pour tout $0 \le k \le 2$. Alors,*

$$\widehat{S}_n(m, X) = \widehat{L}_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m| + \log\big(b_m(\widehat{\theta}_m)\big) + \frac{\log(2\pi)}{2}\,|m| - \frac{1}{2}\,\log\big(\det\big(-\widehat{F}_n(m)\big)\big) - \log(|\mathcal{M}|) + O(n^{-1}) \quad a.s. \quad (1.3.24)$$

avec $\widehat{F}_n(m) := \partial^2_{\theta^2}\widehat{L}_n(\widehat{\theta}_m)$ et $\pi_m = 1/|\mathcal{M}|$.

Dans l'équation ci-dessus, il est clair que $-2\,\widehat{S}_n(m, X) \simeq -2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m|$ p.s.. Cela donne une légitimité au critère BIC habituel dans le cadre des processus affines causaux puisque :

$$\widehat{m}_{BIC} = \text{argmin}_{m\in\mathcal{M}}\Big\{-2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m|\Big\},$$

et nous observons que $\widehat{m}_{BIC}$ maximise les principaux termes de $\widehat{S}_n(m, X)$.

À partir de la relation (1.3.24), en considérant certains termes de second ordre dans l'approximation asymptotique de $\widehat{S}_n(m, X)$, nous pouvons aussi obtenir le critère de Kashyap (voir Kashyap (1982), Sclove (1987), Bozdogan (1987)), définie pour tout $m \in \mathcal{M}$ par

$$\widehat{KC}(m) := -2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m| + \log\big(\det\big(-\widehat{F}_n(m)\big)\big) \quad \text{et} \quad \widehat{m}_{KC} = \text{argmin}_{m\in\mathcal{M}}\big\{\widehat{KC}(m)\big\}. \quad (1.3.25)$$

Par conséquent, nous pourrions définir un nouveau critère consistant data driven, appelé $\widehat{KC}'$, tel que, pour tout $m \in \mathcal{M}$

$$\widehat{KC}'(m) := -2\,\widehat{L}_n(\widehat{\theta}_m) + \big(\log(n) - \log(2\pi)\big)\,|m| + \log\big(\det\big(-\widehat{F}_n(m)\big)\big) + 2\log\big(|m|\big) \quad \text{et} \quad \widehat{m}_{KC'} = \text{argmin}_{m\in\mathcal{M}}\big\{\widehat{KC}'(m)\big\}. \quad (1.3.26)$$

Corollaire 1.1. *Sous les hypothèses* **A1, A2, A3** *et* **A5**, *nous déduisons de Bardet et al. (2020b), que $\widehat{m}_{BIC}$, $\widehat{m}_{KC}$ et $\widehat{m}_{KC'}$ sont consistants, i.e.*

$$\widehat{m}_{BIC} \xrightarrow[n\to\infty]{\mathcal{P}} m^*, \quad \widehat{m}_{KC} \xrightarrow[n\to\infty]{\mathcal{P}} m^* \quad et \quad \widehat{m}_{KC'} \xrightarrow[n\to\infty]{\mathcal{P}} m^*.$$

1.3.3.d Quelques exemples de calcul du terme trace dans la penalité idéale

Pour un modèle m vérifiant $\theta_m^* \in \Theta_m$

1/ Si (ξ_t) est un bruit blanc gaussien, alors $\mu_4 = 3$ et d'après Bardet and Wintenberger (2009), l'on a

$$G(\theta_m^*) = -F(\theta_m^*) \quad \Longrightarrow \quad -2\,\mathrm{Trace}\big(\big(F(\theta_m^*)\big)^{-1} G(\theta_m^*)\big) = 2\,\mathrm{Trace}\big(I_{|m|}\big) = 2\,|m|;$$

2/ Si le paramètre θ identifiant un modèle affine causal $X_t = M_\theta^t\,\xi_t + f_\theta^t$ peut s'écrire $\theta = (\theta_1, \theta_2)'$ avec $f_\theta^t = \widetilde{f}_{\theta_1}^t$ et $M_\theta^t = \widetilde{M}_{\theta_2}^t$. Soit p_1, p_2 tels que $p_1 = |\theta_1|$, $p_2 = |\theta_2|$ et $|m| = p_1 + p_2$. Alors nous obtenons

$$-2\,\mathrm{Trace}\Big(\big(F(\theta_m^*)\big)^{-1} G(\theta_m^*)\Big) = 2\,p_1 + (\mu_4 - 1)\,p_2. \tag{1.3.27}$$

Au Chapitre 4, l'on trouvera l'exemple de calcul dans un cas complexe (AR-ARCH).

1.3.4 Sélection de modèles data driven pour la prédiction d'un processus linéaire autorégressif

A la suite de notre troisième contribution sur l'efficience asymptotique, une question naturelle qui se pose est celle de la construction de critères de sélection qui atteindront cette propriété d'efficience mais à n fixé. Au vu de la compléxité de notre modèle général, nous nous sommes restreints au processus AR(∞).

Supposons avoir observé $(X_1, X_2, \ldots, X_n)$ trajectoire du processus (X_t) vérifiant

$$X_t = \sum_{k=1}^{+\infty} \theta_k^*\,X_{t-k} + \sigma\,\xi_t \quad \text{pour tout } t \in \mathbb{Z}. \tag{1.3.28}$$

L'on aimerait prédire X_{n+1} par sélection de modèles. Ainsi, soit S_m (souvent noté m) un modèle qui est un ensemble de fonctions linéaires f de $\mathbb{R}^{D_m}$ dans $\mathbb{R}$ telle que

$$f(x_1, x_2, \ldots, x_{D_m}) = \sum_{i=1}^{D_m} \theta_i\,x_i, \tag{1.3.29}$$

avec $\theta = (\theta_1, \ldots, \theta_{D_m}) \in \Theta_m$ et Θ_m un compact de $\mathbb{R}^{D_m}$.

Etant donné un prédicteur $f_\theta \in S_m$, sa qualité de prédiction est mesurée par le risque quadratique

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

où $f_\theta^n = f_\theta(X_{n-1}, \ldots, X_{n-D_m})$. Le prédicteur de Bayes qui minimise $R(\theta)$ sur l'ensemble de tous les prédicteurs est clairement la fonction inaccessible f_{θ^*}. Introduisons alors l'excès de risque du prédicteur f_θ

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

we will consider that the excess loss is measured on the design points Pour un modèle m, le meilleur prédicteur est $f_{\theta_m^*}$ défini par

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

La version empirique de θ_m^* est

$$\widehat{\theta}_m = \underset{\theta\in\Theta_m}{\operatorname{argmin}}\, \gamma_n(\theta) \quad \text{avec} \quad \gamma_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}(X_t - f_\theta^t)^2. \tag{1.3.30}$$

Considérons $\mathcal{M}_n$ une famille dénombrable de modèles S_m. Le meilleur modèle de la collection $\mathcal{M}_n$ est déterminé par pénalisation du contraste empirique

$$\widehat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}}\, \{C(m)\} \qquad \text{avec} \qquad C(m) := \gamma_n\big(\widehat{\theta}_m\big) + \operatorname{pen}(S_m). \tag{1.3.31}$$

Le but est de trouver la fonction pen de sorte que $\widehat{m}$ vérifie

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \le C_1 \inf_{m\in\mathcal{M}_n}\Big\{\ell(\theta_m^*, \theta^*) + \operatorname{pen}(S_m)\Big\} + \frac{C_2}{n} \tag{1.3.32}$$

où $C_1 = 1+\delta$ avec $\delta > 0$ (et proche de zéro) et $C_2 > 0$ et $\ell(\widehat{\theta}, \theta^*) = \mathbb{E}\Big[\frac{1}{n}\sum_{t=1}^{n}\big(f_{\widehat{\theta}}^t - f_{\theta^*}^t\big)^2\Big]$. C'est-à-dire que le modèle sélectionné $\widehat{m}$ sera suffisamment grand pour réduire son biais, mais pas trop grand pour éviter d'avoir une variance élevée.

1.3.4.a Hypothèses

Pour montrer (1.3.32), nous faisons les hypothèses suivantes.
La toute première n'est rien d'autre que la condition assurant l'existence et la stationnarité de la solution de (1.3.28)

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

La deuxième hypothèse concerne la sous-gaussianité des observations. Elle nous permettra d'appliquer les inégalités exponentielles pour de tels types de variables.
A2 X_t est sous-gaussien avec proxy de variance $\sigma_0^2 > 0$ i.e.

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

Cette hypothèse est vérifiee si l'on considère un bruit gaussien ou borné.

Pour assurer l'inversibilité des matrices Σ_m et $\widehat{\Sigma}_m$ définies par $\Sigma_m = \mathbb{E}\big[\widehat{\Sigma}_m\big]$, $\widehat{\Sigma}_m = \mathbf{M}_m^\top\mathbf{M}_m$ avec $\mathbf{M}_m = \big[X_{i-1}, \ldots, X_{i-D_m}\big]_{i=1}^n$, il est suffisant de supposer que

A3: Pour tout $f_\theta \in S_m$, $< \alpha, \partial_\theta f_\theta >= 0$ *p.s.* $\implies$ $\alpha = 0$.

Nous avons eu besoin de supposer une condition assez faible sur la densité spectrale du processus $(X_t)_{t\in\mathbb{Z}}$. Cette densité $g : [-\pi, \pi[\longrightarrow \mathbb{C}$ est définie ainsi qu'il suit

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

où $r(h) := \mathbb{E}[X_t\, X_{t+h}]$, est la fonction de covariance. La fonction g existe en vertue de **A1** avec $|\theta_t^*| = O(t^{-\gamma})$ où $\gamma \ge 1$.

A4: Il existe une constante $a > 0$ telle que $\inf\limits_{-\pi\le\lambda<\pi} g(\lambda) \ge a$.

Pour des raisons techniques, nous supposerons que la dimension du plus grand modèle K_n de $\mathcal{M}_n$ est de la forme

1.3.4.b Résultats théoriques

La proposition suivante, qui est une conséquence du Théorème 3.1 de Doukhan and Wintenberger (2008), établit un lien entre les coefficients de mélange τ du processus $(X_t)_{t\in\mathbb{Z}}$ et les coefficients θ_i^* du modèle (1.3.28).

Proposition 1.6. *Supposons que* ***A1*** *tienne avec* $|\theta_t^*| = O(t^{-\gamma})$ *et* $\gamma > 1$*, alors il existe une solution* τ*-dependante stationaire de* (1.3.28) *et une constante* $C_\tau > 0$ *telle que pour tout* $r > 0$

$$\tau_{X,\infty}^{(2)}(r) \leq C_\tau \left(\frac{\log r}{r}\right)^{\gamma-1} \tag{1.3.33}$$

Le résultat intermédiaire suivant nous donne une borne inférieure et supérieure uniformes pour la norme spectrale des matrices Σ_m.

Proposition 1.7. *Supposons que* ***A1*** *tienne avec* $|\theta_t^*| = O(t^{-\gamma})$ *et* $\gamma \geq 2$ *, alors pour tout* $m \in \mathcal{M}_n$

$$\left\|\Sigma_m\right\|_{\text{op}} \leq \pi^{-1} \sum_{i=0}^{\infty} \left|\mathbb{E}[X_0\, X_i]\right| < \infty. \tag{1.3.34}$$

Si ***A3-A4*** *sont satisfaites par ailleurs, alors*

$$\left\|\Sigma_m^{-1}\right\|_{\text{op}} \leq 1/a. \tag{1.3.35}$$

Pour prouver (1.3.32) en présence de la dépendance, il est courant de considérer l'ensemble

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

où $\|F_\theta\|_\mu^2 := \frac{1}{n}\,\mathbb{E}\Big[\sum_{t=1}^n (f_\theta^t)^2\Big] = \int (f_\theta^1)^2 d\mu$ et $\|F_\theta\|_n^2 = n^{-1}\sum_{t=1}^n (f_\theta^t)^2$, $F_\theta = (f_\theta^1, \ldots, f_\theta^n)^\top$.

Sur Ω_n, on a une relation entre la norme empirique $\|.\|_n$ et la norme $\mathbb{L}_2$ (voir par exemple Baraud et al. (2001b), Comte and Genon-Catalot (2020)) et cela est très pratique pour prouver (1.3.32).

La proposition suivante montre que Ω_n tient avec une grande probabilité.

Proposition 1.8. *Sous les hypothèses* ***A1*** $-$ ***A4*** *et si* $|\theta_t^*| = O(t^{-\gamma})$ *avec* $\gamma \geq 8$ *, nous avons*

$$\mathbb{P}(\Omega_n^c) \leq \frac{c_0}{n^3}, \tag{1.3.36}$$

où $c_0 > 0$.

Nous sommes maintenant en mesure d'énoncer le résultat principal de notre quatrième contribution.

Théorème 1.10. *Considérons les observations* $(X_1, \ldots, X_n)$ *d'une solution du processus* (1.3.28) *satisfaisant* ***A1*** *avec* $|\theta_t^*| = O(t^{-\gamma})$ *où* $\gamma \geq 8$ *et vérifiant aussi* ***A2*** *et* ***A4****. Soit* $\mathcal{M}_n$ *une famille dénombrable de modèles* S_m *satisfaisant* ***A3****. Pour* $x > 0$*, considérons la fonction pen:* $\mathcal{M}_n \to \mathbb{R}^+$ *telle que*

$$\text{pen}(S_m) \geq 8\, x\, \sigma^2\, \frac{D_m}{n}. \tag{1.3.37}$$

Alors, l'estimateur $\widehat{\theta}_{\widehat{m}}$ *avec* $\widehat{m}$ *donnée en* (1.3.31)*, satisfait*

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

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

Ce théorème vient clore la liste des résultats principaux obtenus tout au long de cette thèse. Dans la suite, nous allons présenter de manière détaillée les quatre chapitres associés aux contributions résumées dans ce chapitre.

2

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

The content of this chapter was taken from the article: J-M BARDET, K. KAMILA, W. KENGNE, "*Consistent Model Selection Criteria and Goodness-of-fit Test for Affine Causal Processes*", Electronic journal of statistics , Shaker Heights, OH : Institute of Mathematical Statistics, 2020, 14 (1), pp.2009-2052.

Abstract
This chapter 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.

2.1 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 (Stone (1974), Allen (1974)) 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 Mallows (1973) and Akaike (1973). Although it is likely that these ideas have already existed in other contexts such as subset selection by Beale et al. (1967), Hocking and Leslie (1967), and ridge regression by Hoerl and Kennard (1970).
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, Schwarz (1978) 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 Hoerl and Kennard (1970), the $\mathbb{L}^1$ norm used by Tibshirani (1996) that provides the LASSO procedure and the elastic-net that mixes the $\mathbb{L}^1$ and $\mathbb{L}^2$ norms Zou and Hastie (2005).
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:

$$X_t = M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t + f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \text{ for any } t \in \mathbb{Z}; \tag{2.1.1}$$

where $(\xi_t)_{t\in\mathbb{Z}}$ 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 Ding et al. (1993) and Ling and McAleer (2003)) or APARCH(δ, p, q) processes (see Ding et al. (1993)) belongs to $\mathcal{AC}(M, f)$. The existence of stationary and ergodic solutions of this class has been studied in Doukhan and Wintenberger (2008) and Bardet and Wintenberger (2009).

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 McQuarrie and Tsai (1998) 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 Hannan (1980) or Tsay (1984) ; and the asymptotic efficiency of the AIC has been proved, see, among others, Shibata (1980), Hurvich and Tsai (1989) for a corrected version of AIC for small samples, Ing and Wei (2005), Ing (2007), Ing et al. (2012) for the case of infinite order autoregressive model. Shi and Tsai (2002) propose the (consistent) residual information criteria (RIC) for regression model (including regression models with ARMA errors) selection. In the framework of nonlinear threshold models, Kapetanios (2001) proved consistency results of a large class of information criteria, whereas Gao and Tong (2004) 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 Rao et al. (2001), 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 Ding et al. (2018).

The adaptive lasso, introduced by Zou (2006) for variable selection in linear regression models has been extended by Ren and Zhang (2010) to vector autoregressive models, Kock (2016) carried out this procedure in stationary and nonstationary autoregressive models; the oracle efficiency is established. Lerasle (2011) considers model selection for density estimation under mixing conditions and derived oracle inequalities of the slope heuristic procedure (Birgé and Massart (2007a) or Arlot and Massart (2009)) ; whereas Alquier and Wintenberger (2012) develop oracle inequalities for model selection for weakly dependent time series forecasting. Recently, Shao and Yang (2017) have considered the model selection for ARMA time series with trend, and proved the consistency of BIC for the detrended residual sequence, while Arkoun et al. (2019) developed oracle inequalities of sequential model selection method for nonparametric autoregression. Hsu et al. (2019a) 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 (5.1.1), 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 2.2). 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 Li and Mak (1994) under the Gaussian assumption on the observations, whereas Ling and Li (1997) 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. Duchesne and Francq (2008) 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 (5.1.1). 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 also considered a data-driven choice of the penalty obtained from the slope heuristic procedure (see for instance Arlot and Massart (2009)) 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 the Portetemanteau test are overall satisfactory.

The paper is organized as follows. Some definitions, notations and assumptions are described in Section 2.2. The consistency of the criteria and the asymptotic normality of the post-model-selection estimator are studied in Section 2.3. In Section 2.4, 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 2.5. Finally, numerical results are presented in Section 2.6 and Section 2.7 contains the proofs.

2.2 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.

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

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

This approach as semi-parametric estimation has been successively introduced for GARCH(p,q) processes in Jeantheau (1998) where its consistency is also proved, and the asymptotic normality of this estimator has been established in Berkes et al. (2003) and Francq and Zakoïan (2004). In Bardet and Wintenberger (2009), those results have been extended to affine causal processes, and an extension to Laplacian QMLE has been also proposed in Bardet et al. (2017).
The Gaussian QMLE is derived from the conditional (with respect to the filtration $\sigma\{(X_t)_{t\leq 0}\}$) log-likelihood of $(X_1,\dots,X_n)$ 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:

$$L_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} q_t(\theta)\ ,\ \text{with } q_t(\theta) := \frac{(X_t - f_\theta^t)^2}{H_\theta^t} + \log(H_\theta^t) \tag{2.2.1}$$

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 $H_\theta^t = \left(M_\theta^t\right)^2$. 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 (2.2.3) 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

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

$$\text{where} \begin{cases} \widehat{f}_\theta^t & := & f_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{M}_\theta^t & := & M_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{H}_\theta^t & := & (\widehat{M}_\theta^t)^2 \end{cases} \tag{2.2.2}$$

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

$$\widehat{\theta}(m) = \underset{\theta \in \Theta(m)}{\operatorname{argmax}}\, \widehat{L}_n(\theta). \tag{2.2.3}$$

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

$$\widehat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}}\, \widehat{C}(m) \quad \text{with} \quad \widehat{C}(m) = -2\widehat{L}_n\big(\widehat{\theta}(m)\big) + |m|\, \kappa_n, \tag{2.2.4}$$

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 $\{1, \dots, d\}$, 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.*

$$\mathbb{P}(\widehat{m} = m^*) \underset{n \to \infty}{\longrightarrow} 1; \tag{2.2.5}$$

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^*.

2.2.2 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 $\|g\|_\Theta = \sup_{\theta \in \Theta}\big\{\|g(\theta)\|\big\}$.

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 $(X_1, \dots, X_n)$ 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 2.1. *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 2.7. 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.*

$$\left\{ \begin{array}{ll} AR & \left\{ \begin{array}{l} M^{(1)}_{\theta_1}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sigma \\ f^{(1)}_{\theta_1}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} \end{array} \right. \\ \\ ARCH & \left\{ \begin{array}{l} M^{(2)}_{\theta_2}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sqrt{a_0 + a_1 X^2_{t-1} + \cdots + a_q X^2_{t-q}} \\ f^{(2)}_{\theta_2}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = 0 \end{array} \right. \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} M_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \sqrt{\theta_0 + \theta_1 X^2_{t-1} + \cdots + \theta_q X^2_{t-q}} \\ f_\theta\big((X_{t-i})_{i\in\mathbb{N}^*}\big) = \theta_{q+1} X_{t-1} + \cdots + \theta_{q+p} X_{t-p} \end{array} \right. .$$

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

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

$$X_t = M_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\, \xi_t + f_{\theta^*}\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \quad \text{for any } t \in \mathbb{Z}. \tag{2.2.6}$$

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

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

Example: From the previous example, we can consider:

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

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

2.2.3 The special case of NLARCH(∞) processes

As in Bardet and Wintenberger (2009), in the special case of NLARCH(∞) processes, including for instance GARCH(p,q) or ARCH(∞) processes, a particular treatment can be realized for obtaining sharper results than using the previous framework. In such case, define the class:

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

$$X_t = \xi_t \sqrt{\widetilde{H}_\theta\big((X_{t-i}^2)_{i\in\mathbb{N}^*}\big)} \quad \text{for any } t\in\mathbb{Z}. \tag{2.2.7}$$

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

$$\widetilde{\Theta}(r) = \Big\{\theta\in\mathbb{R}^d,\ A(\widetilde{H}_\theta, \{\theta\}) \text{ holds with } \big(\|\xi_0\|_r\big)^2 \sum_{k=1}^{\infty} \alpha_k(\widetilde{H}_\theta, \{\theta\}) < 1\Big\}. \tag{2.2.8}$$

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

2.3 Asymptotic results

2.3.1 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 Doukhan and Wintenberger (2008), is written in terms of Lipschitz coefficients of both these functions. Hence, for $\Psi_\theta = f_\theta$ or M_θ, define:

Assumption $\mathbf{A}(\Psi_\theta, \Theta)$: *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:

$$\Theta(r) = \Big\{\theta\in\mathbb{R}^d,\ A(f_\theta,\{\theta\}) \text{ and } A(M_\theta,\{\theta\}) \text{ hold with } \sum_{k=1}^{\infty}\alpha_k(f_\theta,\{\theta\}) + \|\xi_0\|_r \sum_{k=1}^{\infty}\alpha_k(M_\theta,\{\theta\}) < 1\Big\}. \tag{2.3.1}$$

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 Doukhan and Wintenberger (2008) and Bardet and Wintenberger (2009)).

Secondly, note that the definitions of the conditional log-likelihood (2.2.1) and quasi log-likelihood (2.2.2) 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* $\inf_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ *for all* $x \in \mathbb{R}^\infty$.

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

Assumption Id(Θ): *For all* $\theta, \theta' \in \Theta$, $(f^0_\theta = f^0_{\theta'}$ *and* $M^0_\theta = M^0_{\theta'})$ *a.s.* $\implies \theta = \theta'$.

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 Bardet and Wintenberger (2009), Straumann and Mikosch (2006), White (1982)). 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 $\theta \in \Theta$, $\big(\sum_{i=1}^d \beta_i \frac{\partial f^0_\theta}{\partial\theta^{(i)}} = 0 \implies \forall i = 1,\dots,d,\ \beta_i = 0\ a.s\big)$ *or* $\big(\sum_{i=1}^d \beta_i \frac{\partial H^0_\theta}{\partial\theta^{(i)}} = 0 \implies \forall i = 1,\dots,d,\ \beta_i = 0\ a.s\big)$.

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

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

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

Remark 2.1. *These conditions on* $(\kappa_n)_{n\in\mathbb{N}}$ *have been deduced from conditions for strong law of large numbers obtained in Kounias and Weng (1969) 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. *Geometric case:* $\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\partial_\theta f_\theta,\Theta)+\alpha_j(\partial_\theta M_\theta,\Theta) = O(a^j)$ *with* $0 \leq a < 1$, *then any* (κ_n) *such as* $1/\kappa_n = o(1)$ *can be chosen; for instance* $\kappa_n = \log n$ *or* $\log(\log n)$*; this is the case for instance of ARMA, GARCH, APARCH or ARMA-GARCH processes.*

2. *Riemanian case:* $\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\partial_\theta f_\theta,\Theta)+\alpha_j(\partial_\theta M_\theta,\Theta) = O(j^{-\gamma})$ *with* $\gamma > 1$*:*

- *if* $r \geq 3$ *then*
  - *if* $\gamma > 2$ *then any sequence such as* $1/\kappa_n = o(1)$ *can be chosen;*
  - *if* $1 < \gamma < 2$*, any* (κ_n) *such as* $\kappa_n = O(n^\delta)$ *with* $\delta > 2 - \gamma$ *can be chosen.*
- *if* $1 \leq r < 3$
  - *if* $\gamma > (r+3)/r$ *then any sequence such as* $1/\kappa_n = o(1)$ *can be chosen;*
  - *if* $1 < \gamma < (r+3)/r$ *then any* (κ_n) *such as* $\kappa_n = n^\delta$ *with* $\delta > (r+3)/r - \gamma$ *can be chosen.*

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

2.3.2 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 2.1. *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*

$$\mathbb{P}(\widehat{m} = m^*) \underset{n\to\infty}{\longrightarrow} 1 \quad \textit{and} \quad \widehat{\theta}(\widehat{m}) \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} \theta^*. \tag{2.3.2}$$

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

Theorem 2.2. *Under the assumptions of Theorem 2.1 and if* $\theta^* \in \mathring{\Theta}$ *and Var*(Θ) *holds, then*

$$\sqrt{n}\left(\big(\widehat{\theta}(\widehat{m})\big)_i - (\theta^*)_i\right)_{i\in m^*} \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}_{|m^*|}\big(0 \;\; F(\theta^*, m^*)^{-1} G(\theta^*, m^*) F(\theta^*, m^*)^{-1}\big) \tag{2.3.3}$$

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

Remark 2.2. *In Remark 2.1, we detailed some situations where the assumption* $K(\Theta)$ *(or* $\widetilde{K}(\Theta)$*) holds, which leads to the results of Theorem 2.1 and 2.2. 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.*

2.4 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').

2.4.1 AR(∞) 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:

$$X_t = \sum_{k\geq 1} \psi_k(\theta^*) X_{t-k} + \sigma\, \xi_t \quad \text{for any } t \in \mathbb{Z}, \tag{2.4.1}$$

where $(\xi_t)_t$ admits 4-order moments, and $\theta^* \in \Theta \subset \Theta(4)$, the set of $\theta \in \mathbb{R}^d$ such that $\sum_{k\geq 1} \|\psi_k(\theta)\|_\Theta < 1$ and $\sigma > 0$. This process corresponds to (2.2.6) 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 $Id(\Theta)$, $\mathrm{Var}(\Theta)$ 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 2.1,

- 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 (2.2) holds if in addition $\theta^* \in \overset{\circ}{\Theta}$;
- if $1 < \gamma < 2$, $\kappa_n = O(n^\delta)$ with $\delta > 2 - \gamma$ has to be chosen (and we cannot insure the consistency of $\widehat{m}$ in case of classical BIC penalization).

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

2.4.2 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 :

$$X_t = \Big(\psi_0(\theta^*) + \sum_{k=1}^{\infty} \psi_k(\theta^*) X_{t-k}^2\Big)^{1/2} \xi_t \quad \text{for any } t \in \mathbb{Z}, \tag{2.4.2}$$

where $\mathbb{E}\big[\xi_0^4\big] < \infty$, and $\theta^* \in \Theta \subset \widetilde{\Theta}(4)$, the set of $\theta \in \mathbb{R}^d$ such that $\sum_{k\geq 1} \|\psi_k(\theta)\|_\Theta < 1$. This process corresponds to (2.2.6) 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 $\sum_k \|\partial^2_\theta \psi_k(\theta)\|_\Theta < \infty$ and $\|\psi_k(\theta)\|_\Theta + \|\partial_\theta \psi_k(\theta)\|_\Theta = O(k^{-\gamma})$ with $\gamma > 1$. From Remark 2.1,

- 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 (2.2) 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.

2.4.3 APARCH(δ, p, q) models

For $\delta \geq 1$ and from Ding et al. (1993), $(X_t)_{t\in\mathbb{Z}}$ is an APARCH(δ, p, q) process with $p, q \geq 0$ if:

$$\begin{cases} X_t = \sigma_t\, \xi_t \\ (\sigma_t)^\delta = \omega + \sum_{i=1}^p \alpha_i (|X_{t-i}| - \gamma_i X_{t-i})^\delta + \sum_{j=1}^q \beta_j (\sigma_{t-j})^\delta \quad \text{for any } t \in \mathbb{Z}, \end{cases} \tag{2.4.3}$$

where $\omega > 0$, $-1 < \gamma_i < 1$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for $1 \leq i \leq p$ and $1 \leq j \leq q$, $\alpha_p > 0$, $\beta_q > 0$ and $\sum_{j=1}^q \beta_j < 1$. From Bardet et al. (2017), with $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \gamma_1, \ldots, \gamma_p, \beta_1, \ldots, \beta_p)'$, the conditional variance σ_t can be rewritten as follows

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

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

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

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

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

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

With all the previous conditions, assumptions D(Θ), Id(Θ), Var(Θ) are satisfied. Moreover, since the Lipschitz coefficients decrease exponentially fast, K(Θ) is satisfied when $\kappa_n \to \infty$. Therefore, the consistency Theorem (2.1) and the Theorem (2.2) of the estimator of the chosen model are satisfied when $r = 4$ and $\kappa_n \to \infty$ (for instance with the typical BIC penalty $\kappa_n = \log n$).

2.4.4 ARMA(p,q)-GARCH(p',q') models

From Ding et al. (1993) and Ling and McAleer (2003), we define $(X_t)_{t\in\mathbb{Z}}$ as an (invertible) ARMA(p,q)-GARCH(p',q') process with $p,q,p',q' \geq 0$ if:

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

where

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

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

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

Therefore, if $(a_1, \ldots, d_{q'}) \in \Theta_{p,q,p',q'}$, $(\varepsilon_t)_t$ is a stationary GARCH(p',q') process and $(X_t)_t$ is a stationary weak invertible ARMA(p,q) process.
Moreover, following Lemma 2.1. of Bardet et al. (2017), we know that a stationary ARMA(p,q)-GARCH(p',q') process is a stationary affine causal process with functions f_θ and M_θ satisfying the Assumption A(f_θ, Θ) and A(M_θ, Θ) with Lipschitzian coefficients decreasing exponentially fast, as well as their derivatives. Finally, if Θ is a bounded subset of $\Theta_{p,q,p',q'}$, then assumptions D(Θ), Id(Θ) and Var(Θ) are automatically satisfied.

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

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

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

With Θ a bounded subset of $\Theta_{p_{\max},q_{\max},p'_{\max},q'_{\max}}$, all the previous assumptions D(Θ), Id(Θ), Var(Θ) are satisfied and K(Θ) is also satisfied as soon as $\kappa_n \to \infty$. As a consequence, in this framework the consistency Theorem (2.1) and the Theorem (2.2) of the estimator of the chosen model are satisfied when $r = 4$ and $\kappa_n \to \infty$ (for instance with the typical BIC penalty $\kappa_n = \log n$).

2.5 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 Li and Mak (1994), Ling and Li (1997)) and the absolute residual (or usual residuals) correlogram (see for instance Li (1992), Duchesne and Francq (2008), Li and Li (2008)), 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 (Li and Mak (1994), Ling and Li (1997)) 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

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

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 2.3. *Under the assumptions of Theorem 2.2, with also*

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

then,

1. *With $V(\theta^*, m^*)$ defined in (2.7.39), it holds that*

$$\sqrt{n}\,\widehat{\rho}(m^*) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\big(0\,,\, V(\theta^*, m^*)\big). \tag{2.5.1}$$

2. *With* $\widehat{Q}_K(m^*) := n\,\widehat{\rho}(m^*)'\big(V(\widehat{\theta}(m^*), m^*)\big)^{-1}\widehat{\rho}(m^*)$, *we have*

$$\widehat{Q}_K(m^*) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(K). \tag{2.5.2}$$

3. *The previous points 1. and 2. also hold when* m^* *is replaced by* $\widehat{m}$.

Using the Theorem 2.3, we can asymptotically test:

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

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 2.3. 1. *In practice the constant* μ_4 *and the columns of the matrix* $J_K(m^*)$ *(see* (2.7.35)*) 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* $\big(\widehat{J}_K(\widehat{\theta}(\widehat{m})\big)_{.,k} = \frac{1}{n}\sum_{t=1}^{n-k}[(\widehat{e}_t(\widehat{m}))^2 - 1]\partial_\theta \log\big(M^{t+k}_{\widehat{\theta}(\widehat{m})}\big)$.

2. *For* $AR(\infty)$ *models (and then for causal invertible* $ARMA(p,q)$*), since* $M_\theta = \sigma$ *as we have seen in Sub-section 2.4.1, we deduce from* (2.7.39) *that* $V(\theta^*, m^*) = I_K$ *as* $J_K(m^*) = 0$. *Hence, in such a case, we simply obtained:*

$$\widehat{Q}_K(\widehat{m}) = n\left\|\widehat{\rho}(\widehat{m})\right\|^2 \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(K). \tag{2.5.3}$$

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 $K \in \mathbb{N}^*$.

2.6 Numerical Results

This section features some simulation experiments that are performed to assess the usefulness of the asymptotic results obtained in Section 2.3. 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.

2.6.1 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: $X_t = \sigma_t \xi_t$, with $\sigma_t^2 = 0.2 + 0.3X_{t-1}^2 + 0.5\sigma_{t-1}^2$.

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

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

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 2.1. More precisely, for each penalty $(\log n,\ \sqrt{n})$ 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).

Table 2.1: Percentage of selected order based on 1000 replications depending on sample's length for Model 1, 2, 3 and 4 respectively.

	Sample length n	100		500		1000		2000	
	Penalty	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$
Model 1	Wrong	21.4	32.3	1.7	0.8	0.8	0.1	0.2	0
	True	74.2	67.6	97.2	99.2	98.2	99.9	99.2	100
	Overfitted	4.4	0.1	1.1	0	1.0	0	0.6	0
Model 2	Wrong	30.4	57.7	4.8	4.2	0.7	0.3	0.4	0
	True	64.1	42.1	93.6	95.8	98.2	99.7	99.2	100
	Overfitted	5.5	0.2	1.6	0	1.1	0	0.4	0
Model 3	Wrong	76.1	90.8	27.3	67.1	14.0	41.5	4.6	12.0
	True	23.8	9.2	72.7	32.9	85.9	58.5	95.4	88.0
	Overfitted	0.1	0	0	0	0.1	0	0	0
Model 4	Wrong	83.8	94.3	22.1	61.5	5.8	31.3	1.8	6.2
	True	15.9	5.7	77.5	38.5	93.2	68.7	98.0	93.8
	Overfitted	0.3	0	0.4	0	1.0	0	0.2	0

From these results, it is clear that the consistency of our model selection procedure is numerically convincing, which is in accordance with Theorem 2.1, where both 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.
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 2.2 shows the empirical size and empirical power of this test. We call by empirical size, the percentage of falsely rejecting the null hypothesis H_0. On the other hand, the empirical power represents the percentage of rejection of H_0 when we arbitrary chose a false model, which is a AR(3) process $X_t = 0.2X_{t-1} + 0.2X_{t-2} + 0.4X_{t-1} + \xi_t$ for Model 1 and 2, and a ARCH(3) process $X_t = \xi_t \sqrt{0.4 + 0.2X_{t-1}^2 + 0.2X_{t-2}^2 + 0.2X_{t-3}^2}$ for Model 3 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 Li and Mak (1994) 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, Tse and Zuo (1997) suggested that $K = p + q + 1$ may be an appropriate choice for the GARCH(p, q) family. Thus, in our tests, we consider $K = 3$ and $K = 6$ so that the rejection is based on the upper 5th percentile of the $\chi^2(3)$ distribution on the one hand and $\chi^2(6)$ on the other hand.

Table 2.2: 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

Once again, the results of Table 2.2 numerically confirms the asymptotic results of Theorem 2.3. Remark that the test is more powerful by using values of K not too large as mentioned above especially for small samples.

2.6.2 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:

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

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

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

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

As a consequence, $1024 = 2^{10}$ candidate models are considered and Table 2.3 presents the results of the selection procedure.

Table 2.3: Percentage of selected model based on 1000 replications depending on sample's length for Model 4

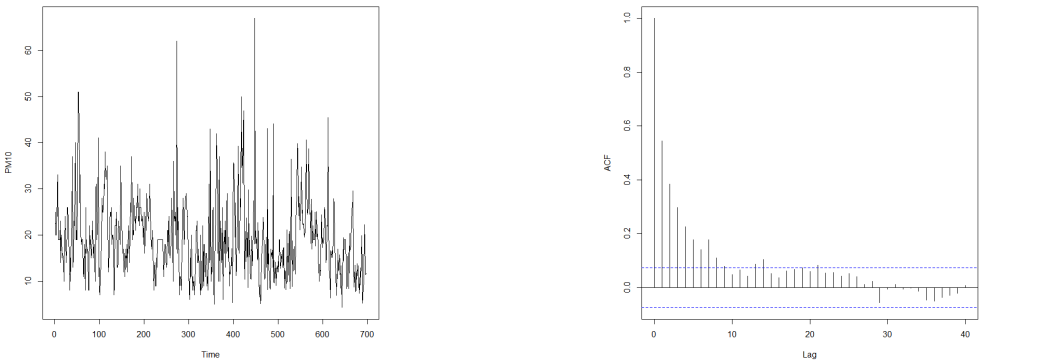
Sample length	100		500		1000		2000	
	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$	$\log n$	$\sqrt{n}$
True model	70.4	67.3	90.0	100	93.2	100	95.3	100
Overfitted	25.0	1.6	10	0	6.8	0	4.7	0
False model	4.6	31.1	0	0	0	0	0	0

We deduce that the consistency of our model selection procedure is also numerically convincing in this case of exhaustive model selection, which is in accordance with Theorem 2.1.

2.6.3 Application to real data

2.6.3.a 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 the increasing number of human activities, the air is being degraded by a wide variety of pollutants, including PM. PM stands for particulate matter government (2017): 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 2.1a). We are going to use our model selection criteria to identify the "best" model for this time series.

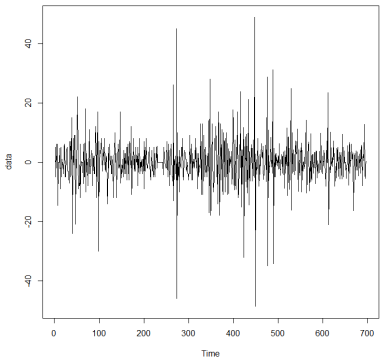


(a) Time plot of PM10 levels.

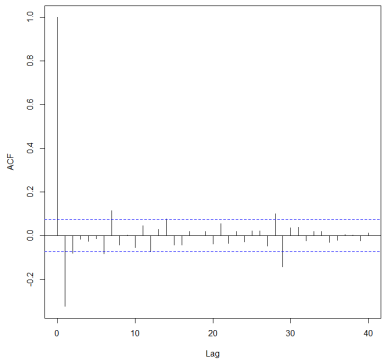
(b) The sample autocorrelation function for the Marseille PM10 showing the bounds $\pm 1.96/\sqrt{n}$.

Figure 2.1: The Marseille PM10 levels (January 1st, 2018 to November 30, 2019).

An inspection of the Figure 2.1 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 2.2). Therefore, we use the same family $\mathcal{M}$ already considered in Subsection 2.6.1 that provides us 66 candidate models. For each model, we compute the criterion (2.2.4) with $\kappa_n = \log(n)$ and $\kappa_n = \sqrt{n}$ The selection results and also the goodness-of-fit of the selected model are featured in the Table 2.4.



(a) PM10 levels.



(b) The sample autocorrelation function for the Marseille PM10 showing the bounds $\pm 1.96/\sqrt{n}$.

Figure 2.2: Elimination of trend and seasonality in Marseille PM10 levels (January 1st, 2018 to November 30, 2019).

Table 2.4: Summary of the results of the model selection and goodness-of-fit analysis on PM10.

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

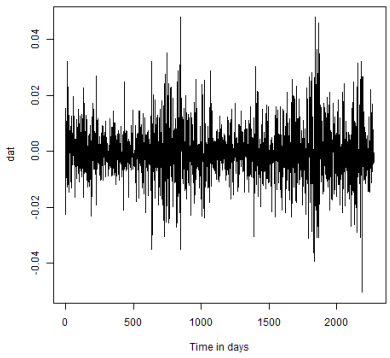
This table shows that all p-values are greater than 0.05, and then none of the test statistics leads us to reject the null hypothesis at this level even though the case of the ARMA$(1,1)$ is somehow limit. The chosen ARMA$(1,2)$ seems to be the more suitable model for PM10 time series.

2.6.3.b Financial index analysis

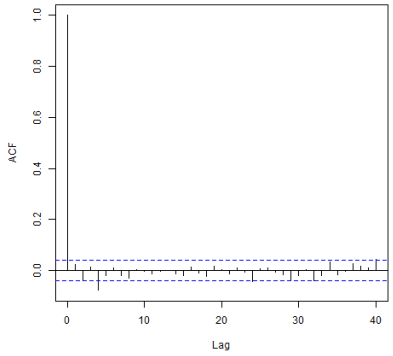
We consider the returns of the daily closing prices of the FTSE 100 index and also the SP 500. They are respectively 2273 and 2264 observations from January 4th, 2010 to December 31st, 2018 for FTSE 100 and SP500. The time plot and the correlograms for the log-returns and squared log-returns are plotted in Figure 2.3. Figures 2.3a and 2.3c exhibit the conditional heteroskedasticity in the log-return time series. Moreover, Figure 2.3b 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 2.3d 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 2.4)
As in the previous illustrative example, the ARMA-GARCH is a plausible family for modeling of the FTSE 100 and SP 500 index. The $\log n$ and $\sqrt{n}$ penalizations have been applied to identify the best order and the goodness-of-fit of the selected model has been tested by the Portmanteau test.

Table 2.5: Summary of the results of the model selection and goodness-of-fit analysis on FTSE index.

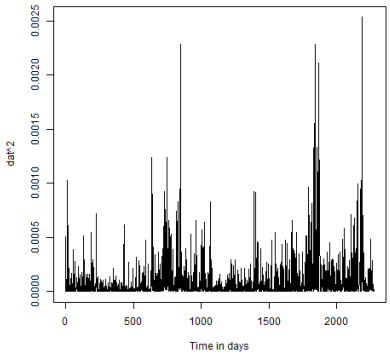
	$\kappa_n = \log(n)$	$\kappa_n = \sqrt{n}$
$\widehat{m}$	GARCH$(1,1)$	GARCH$(1,1)$
$\widehat{Q}_{10}(\widehat{m})$	9.30	9.30
$p - value$	0.50	0.50



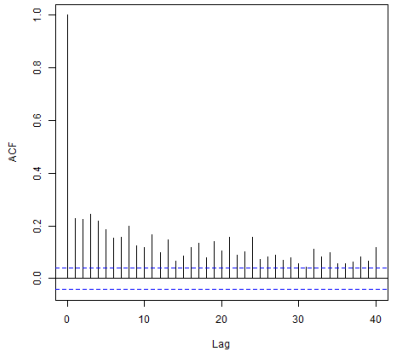
(a) Time plot of log-returns.



(b) Correlograms of log-returns.



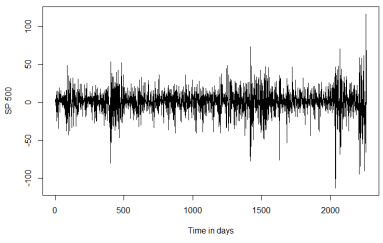
(c) Time plot of squared log-returns.



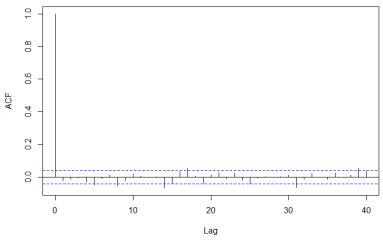
(d) Correlograms of squared log-returns.

Figure 2.3: Daily closing FTSE 100 index (January 4th, 2010 to December 31 st, 2018).

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 are not surprising since the GARCH$(1,1)$ is the reference model and the most commonly used in empirical studies. In addition, Francq and Zakoian (2010) 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.



(a) Time plot of log returns .



(b) The sample autocorrelation function showing the bounds $\pm 1.96/\sqrt{n}$.

Figure 2.4: Daily closing price of SP500 (January 4th, 2010 to December 31 st, 2018).

2.7 Proofs

We start with the proof of the Proposition 2.1.

Proof. For ease of writing, consider only the general case where $f_{\theta_i}^{(i)} = g_{\alpha_i}^{(i)}$ and $M_{\theta_i}^{(i)} = N_{\beta_i}^{(i)}$ 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 $g_{\alpha_1}^{(1)} = h_\alpha + \ell_{\alpha_1'}^{(1)}$, $f_{\alpha_2}^{(2)} = h_\alpha + \ell_{\alpha_2'}^{(2)}$ with $\alpha_1 = {}^t(\alpha, \alpha_1')$ and $\alpha_2 = {}^t(\alpha, \alpha_2')$ and $\ell_0^{(i)} = 0$.
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 $N_{\beta_1}^{(1)} = R_\beta + m_{\beta_1'}^{(1)}$, $N_{\beta_2}^{(2)} = R_\beta + m_{\beta_2'}^{(2)}$ with $\beta_1 = {}^t(\beta, \beta_1')$ and $\beta_2 = {}^t(\beta, \beta_2')$ and $m_0^{(i)} = 0$.
Consider now $\theta = {}^t(\alpha, \alpha_1', \alpha_2', \beta, \beta_1', \beta_2') \in \mathbb{R}^d$ (and therefore $\max(d_1, d_2) \leq d \leq d_1 + d_2$), $f_\theta = h_\alpha + \ell_{\alpha_1'}^{(1)} + \ell_{\alpha_2'}^{(2)}$ and $M_\theta = R_\beta + m_{\beta_1'}^{(1)} + m_{\beta_2'}^{(2)}$. Then if $X \in \mathcal{AC}\big(M_\theta, f_\theta\big)$, for any $t \in \mathbb{Z}$,

$$X_t = \big(R_\beta((X_{t-k})_{k\geq 1}) + m_{\beta_1'}^{(1)}((X_{t-k})_{k\geq 1}) + m_{\beta_2'}^{(2)}((X_{t-k})_{k\geq 1})\big)\,\xi_t + \big(h_\alpha((X_{t-k})_{k\geq 1}) + \ell_{\alpha_1'}^{(1)}((X_{t-k})_{k\geq 1}) + \ell_{\alpha_2'}^{(2)}((X_{t-k})_{k\geq 1})\big).$$

Then, for $\alpha_2' = \beta_2' = 0$, $X \in \mathcal{AC}\big(M_{\theta_1}^{(1)}, f_{\theta_1}^{(1)}\big)$ and for $\alpha_1' = \beta_1' = 0$, $X \in \mathcal{AC}\big(M_{\theta_2}^{(2)}, f_{\theta_2}^{(2)}\big)$. □

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

Lemma 2.1. *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:*

$$\frac{1}{\kappa_n}\big\|\widehat{L}_n(\theta) - L_n(\theta)\big\|_\Theta \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} 0. \tag{2.7.1}$$

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 Kounias and Weng (1969), with $r \leq 3$, (2.7.1) is established when:

$$\sum_{k\geq 1}\Big(\frac{1}{\kappa_k}\Big)^{r/3}\mathbb{E}\big(\|\widehat{q}_k(\theta) - q_k(\theta)\|_\Theta^{r/3}\big) < \infty. \tag{2.7.2}$$

With $r \geq 3$, and under the assumptions, we first recall some results already obtained in Bardet and Wintenberger (2009): for any $t \in \mathbb{Z}$,

- $\mathbb{E}\big[|X_t|^r + \|f_\theta^t\|_\Theta^r + \|\widehat{f}_\theta^t\|_\Theta^r + \|M_\theta^t\|_\Theta^r + \|\widehat{M}_\theta^t\|_\Theta^r + \|H_\theta^t\|_\Theta^{r/2} + \|\widehat{H}_\theta^t\|_\Theta^{r/2}\big] < \infty$ (2.7.3)
- $$\begin{cases} \mathbb{E}\big[\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^r\big] \leq C\left(\sum_{j\geq t}\alpha_j(f_\theta, \Theta)\right)^r \\ \mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r\big] \leq C\left(\sum_{j\geq t}\alpha_j(M_\theta, \Theta)\right)^r \\ \mathbb{E}\big[\|H_\theta^t - \widehat{H}_\theta^t\|_\Theta^{r/2}\big] \leq C\left(\min\left\{\sum_{j\geq t}\alpha_j(M_\theta, \Theta)\ ,\ \sum_{j\geq t}\alpha_j(H_\theta, \Theta)\right\}\right)^{r/2}. \end{cases} \tag{2.7.4}$$

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

$$
\begin{aligned}
\mathbb{E}\big[\|\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta^{r/3}\big] \le C\,\Big(&\mathbb{E}\Big[\big(\|X_t + f_\theta^t\|_\Theta^2 + 1\big)^{r/3}\,\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^{r/3}\Big] \\
&+ \mathbb{E}\Big[\big(2|X_t| + \|f_\theta^t\|_\Theta + \|\widehat{f}_\theta^t\|_\Theta\big)^{r/3}\,\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^{r/3}\Big]\Big). \quad (2.7.5)
\end{aligned}
$$

Then, by Hölder's inequality and (2.7.3) we have:

$$
\begin{aligned}
&\mathbb{E}\Big[\big(\|X_t + f_\theta^t\|_\Theta^2 + 1\big)^{r/3}\,\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^{r/3}\Big] \\
&\le \Big(\mathbb{E}\big[\|X_t + f_\theta^t + 1\|_\Theta^r\big]\Big)^{2/3}\,\Big(\mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r\big]\Big)^{1/3} \le C\,\Big(\mathbb{E}\big[\|M_\theta^t - \widehat{M}_\theta^t\|_\Theta^r\big]\Big)^{1/3}. \quad (2.7.6)
\end{aligned}
$$

Again with Hölder's inequality and (2.7.3) ,

$$
\mathbb{E}\big[\big((2|X_t| + \|f_\theta^t\|_\Theta + \|\widehat{f}_\theta^t\|_\Theta)\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta\big)^{r/3}\big] \le C\,\big(\mathbb{E}\big[\|f_\theta^t - \widehat{f}_\theta^t\|_\Theta^r\big]\big)^{1/3}. \quad (2.7.7)
$$

Therefore, from (2.7.6), (2.7.7) and (2.7.4), there exists a constant C such that

$$
\mathbb{E}\big[\|(\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta^{r/3}\big] \le C\,\Big(\sum_{j\ge t}\alpha_j(f_\theta, \Theta) + \sum_{j\ge t}\alpha_j(M_\theta, \Theta)\Big)^{r/3}. \quad (2.7.8)
$$

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 Kounias and Weng (1969), with $r \le 4$, (2.7.1) is established when:

$$
\sum_{k\ge 1}\big(\frac{1}{\kappa_k}\big)^{r/4}\mathbb{E}\big(\|\widehat{q}_k(\theta) - q_k(\theta)\|_\Theta^{r/4}\big) < \infty. \quad (2.7.9)
$$

By proceeding as in the previous case, we deduce

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

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

$$\mathbb{E}\big[\|(\widehat{q}_t(\theta) - q_t(\theta)\|_{\Theta}^{r/4}\big] \leq C\Big(\sum_{j\geq t}\alpha_j(H_\theta,\Theta)\Big)^{r/4}. \tag{2.7.10}$$

□

Lemma 2.2. *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:*

$$\frac{1}{\kappa_n}\Big\|\frac{\partial\widehat{L}_n(\theta)}{\partial\theta} - \frac{\partial L_n(\theta)}{\partial\theta}\Big\|_{\Theta} \xrightarrow[n\to+\infty]{a.s.} 0. \tag{2.7.11}$$

Proof. We will go along similar lines as in the proof of Lemma 2.1. 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 Kounias and Weng (1969), it is sufficient to prove for $r \leq 3$ that

$$\sum_{k\geq 1}(\frac{1}{\kappa_k})^{r/3}\,\mathbb{E}\Big[\Big\|\frac{\partial\widehat{q}_t(\theta)}{\partial\theta_i} - \frac{\partial q_t(\theta)}{\partial\theta_i}\Big\|_{\Theta}^{r/3}\Big] < \infty. \tag{2.7.12}$$

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,

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

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

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

Thus,

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

Using again the results of Bardet and Wintenberger (2009), we know that:

- $$\mathbb{E}\Big[\Big\|\frac{\partial f^t_\theta}{\partial\theta_i}\Big\|^r_\Theta+\Big\|\frac{\partial\widehat{f}^t_\theta}{\partial\theta_i}\Big\|^r_\Theta+\Big\|\frac{\partial M^t_\theta}{\partial\theta_i}\Big\|^r_\Theta+\Big\|\frac{\partial\widehat{M}^t_\theta}{\partial\theta_i}\Big\|^r_\Theta+\Big\|\frac{\partial H^t_\theta}{\partial\theta_i}\Big\|^{r/2}_\Theta+\Big\|\frac{\partial(H^t_\theta)^{-1}}{\partial\theta_i}\Big\|^r_\Theta\Big]<\infty \quad (2.7.13)$$
- $$\begin{cases}\mathbb{E}\big[\|(H^t_\theta)^{-1}-(\widehat{H}^t_\theta)^{-1}\|^r_\Theta\big]\le C\Big(\sum_{j\ge t}\alpha_j(M_\theta,\Theta)\Big)^r\\ \mathbb{E}\Big[\Big\|\frac{\partial f^t_\theta}{\partial\theta_i}-\frac{\partial\widehat{f}^t_\theta}{\partial\theta_i}\Big\|^r_\Theta\Big]\le C\Big(\sum_{j\ge t}\alpha_j(\partial f_\theta,\Theta)\Big)^r\\ \mathbb{E}\Big[\Big\|\frac{\partial H^t_\theta}{\partial\theta_i}-\frac{\partial\widehat{H}^t_\theta}{\partial\theta_i}\Big\|^{r/2}_\Theta\Big]\le C\Big(\sum_{j\ge t}\big(\alpha_j(M_\theta,\Theta)+\alpha_j(\partial M_\theta,\Theta)\big)\Big)^{r/2}\\ \mathbb{E}\Big[\Big\|\frac{\partial(H^t_\theta)^{-1}}{\partial\theta_i}-\frac{\partial(\widehat{H}^t_\theta)^{-1}}{\partial\theta_i}\Big\|^{r/2}_\Theta\Big]\le C\Big(\sum_{j\ge t}\big(\alpha_j(M_\theta,\Theta)+\alpha_j(\partial M_\theta,\Theta)\big)\Big)^{r/2}\end{cases} \quad (2.7.14)$$

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

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

Using (2.7.13) and (2.7.14), we deduce

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

Therefore,

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

We conclude the proof of (2.7.12) from assumption $K(\Theta)$.

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

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

As a consequence,

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

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 Kounias and Weng (1969). □

Lemma 2.3. *Under the assumptions of Theorem 2.1 and if a model $m\in\mathcal{M}$ is such that $\theta^*\in\Theta(m)$, then:*

$$\frac{1}{\kappa_n}\big|\widehat{L}_n(\widehat{\theta}(m))-\widehat{L}_n\big(\widehat{\theta}(m^*)\big)\big|=o_P(1). \tag{2.7.15}$$

Proof. We have:

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

According to Lemma 2.1, $\frac{1}{\kappa_n}\big\|\widehat{L}_n(\theta)-L_n(\theta)\big\|_\Theta\xrightarrow[n\to+\infty]{a.s.}0$. The proof will be achieved if we can show that

$$\frac{1}{\kappa_n}\big|L_n(\widehat{\theta}(m))-L_n(\theta^*)\big|=o_P(1). \tag{2.7.16}$$

Since

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

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

$$\frac{1}{\kappa_n}\big(L_n(\widehat{\theta}(m)) - L_n(\theta^*)\big) = \frac{1}{\kappa_n}\big(\widehat{\theta}(m) - \theta^*\big)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta} + \frac{1}{2\kappa_n}\big(\widehat{\theta}(m) - \theta^*\big)'\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta^2}\big(\widehat{\theta}(m) - \theta^*\big). \quad (2.7.17)$$

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 Bardet and Wintenberger (2009) we have $\sqrt{n}\big(\widehat{\theta}(m) - \theta^*\big) = O_P(1)$ and $\frac{1}{\sqrt{n}}\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta} = o_P(1)$, it follows that:

$$\frac{1}{\kappa_n}\big(\widehat{\theta}(m) - \theta^*\big)\frac{\partial L_n(\widehat{\theta}(m))}{\partial\theta} = o_P(1). \quad (2.7.18)$$

On the other hand, for the second term of the right hand side of equality (2.7.17), let us note that, we have from Bardet and Wintenberger (2009):

- $\sqrt{n}\left(\widehat{\theta}(m) - \theta^*\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{A}_{\theta^*,m}$ a Gaussian random variable from Theorem 2 of Bardet and Wintenberger (2009).
- $-\frac{2}{n}\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m} \xrightarrow[n\to+\infty]{a.s.} F(\theta^*, m)$ since $\widehat{\theta}(m) \xrightarrow[n\to+\infty]{a.s.} \theta^*$ and using the assumption $\mathrm{Var}(\Theta)$ insuring that the matrix $F(\theta^*, m)$ exists and is definite positive (see Bardet and Wintenberger (2009)).

Hence,

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

We deduce that

$$\big(\widehat{\theta}(m) - \theta^*\big)'\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\big(\widehat{\theta}(m) - \theta^*\big) = O_P(1) \implies \frac{1}{\kappa_n}\big(\widehat{\theta}(m) - \theta^*\big)'\left(\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta_i\partial\theta_j}\right)_{i,j\in m}\big(\widehat{\theta}(m) - \theta^*\big) = o_P(1). \quad (2.7.19)$$

Thus, (2.7.16) follows from (2.7.17), (2.7.18) and (2.7.19); which completes the proof of Lemma 2.3. □

2.7.1 Misspecified model

When a model m is misspecified ($\theta^* \notin \Theta(m)$), we will show that $\mathbb{P}(m^* \not\subseteq \widehat{m}) \underset{n\to\infty}{\longrightarrow} 0$ by following the key idea of similar proof in Sin and White (1996) 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

$$\theta^*(m) := \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\, L(\theta) \quad \text{with} \quad L(\theta) := -\frac{1}{2}\,\mathbb{E}[q_0(\theta)]. \tag{2.7.20}$$

Given that the function L is continuous and the parameter set $\Theta(m)$ is assumed compact, the parameter $\theta^*(m)$ exists and is unique by virtue of the identifiability condition $Id(\Theta)$.

It is worth noting, since $L(\theta)$ has a unique maximum reached at θ^* (see Bardet and Wintenberger (2009)), 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 2.2. *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 $r \geq 2$. Under the assumptions $Id(\Theta)$, $D(\Theta)$ and $K(\Theta)$, it holds*

$$\left\| \frac{1}{n} L_n(\theta) - L(\theta) \right\|_{\Theta(m)} \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} 0 \quad \text{and} \tag{2.7.21}$$

$$\widehat{\theta}(m) \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} \theta^*(m). \tag{2.7.22}$$

Proof. The proof of (2.7.21) follows from a consequence of uniform strong law of large numbers for stationary ergodic sequence (see the proof of Theorem 1 in Bardet and Wintenberger (2009)). The second result holds by applying (2.7.21) and Lemma 2.1. □

2.7.2 Proof of Theorem 2.1

Before diving into the proof, remark first that:

$$\mathbb{P}(\widehat{m} = m^*) = 1 - \mathbb{P}(m^* \subset \widehat{m}) - \mathbb{P}(m^* \not\subseteq \widehat{m}). \tag{2.7.23}$$

As we point out in Subsection 2.2.1, 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:

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

by vertue of Lemma 2.3 and because $|m_0| - |m^*| \ge 1$.
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 $\widehat{m}$ (2.2.4), 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,

$$\begin{aligned}\frac{1}{n}\Big[\widehat{L}_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m_0)\big)\Big] &= \frac{1}{n}\Big[L_n\big(\widehat{\theta}(m^*)\big) - L_n\big(\widehat{\theta}(m_0)\big)\Big] + o_{a.s}(1) \text{ using Lemma 2.1}\\ &= L\big(\widehat{\theta}(m^*)\big) - L\big(\widehat{\theta}(m_0)\big) + o_{a.s}(1) \text{ using Proposition 2.2}\\ &= \big[L\big(\widehat{\theta}(m^*)\big) - L(\theta^*)\big] - \big[L\big(\widehat{\theta}(m_0)\big) - L(\theta^*(m_0))\big]\\ &\qquad + \big[L(\theta^*) - L(\theta^*(m_0))\big] + o_{a.s}(1).\end{aligned}$$

Since L is continuous over Θ, using continuous mapping theorem and the relation (2.7.22) of Proposition 2.2, 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,

$$\frac{1}{n}\Big[\widehat{L}_n\big(\widehat{\theta}(m^*)\big) - \widehat{L}_n\big(\widehat{\theta}(m_0)\big)\Big] = L(\theta^*) - L(\theta^*(m_0)) + o_{a.s}(1). \tag{2.7.24}$$

Now, let us show that that $A(m) := L(\theta^*) - L(\theta^*(m)) > 0$ for $m = m_0$.
Let us denote by $\mathcal{F}_t := \sigma\big(X_{t-1}, X_{t-2}, \cdots\big)$. Using conditional expectation, we obtain

$$L(\theta^*) - L(\theta) = \frac{1}{2}\,\mathbb{E}\Big[\mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big]\Big]. \tag{2.7.25}$$

But,

$$\begin{aligned}\mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big] &= \mathbb{E}\Big[\frac{(X_0 - f^0_\theta)^2}{H^0_\theta} + \log(H^0_\theta) - \frac{(X_0 - f^0_{\theta^*})^2}{H^0_{\theta^*}} - \log(H^0_{\theta^*}) \mid \mathcal{F}_0\Big]\\ &= \log\Big(\frac{H^0_\theta}{H^0_{\theta^*}}\Big) + \frac{\mathbb{E}\big[(X_0 - f^0_\theta)^2 \mid \mathcal{F}_0\big]}{H^0_\theta} - \frac{\mathbb{E}\big[(X_0 - f^0_{\theta^*})^2 \mid \mathcal{F}_0\big]}{H^0_{\theta^*}}\\ &= \log\Big(\frac{H^0_\theta}{H^0_{\theta^*}}\Big) - 1 + \frac{\mathbb{E}\big[(X_0 - f^0_{\theta^*} + f^0_{\theta^*} - f^0_\theta)^2 \mid \mathcal{F}_0\big]}{H^0_\theta}\\ &= \frac{H^0_{\theta^*}}{H^0_\theta} - \log\Big(\frac{H^0_{\theta^*}}{H^0_\theta}\Big) - 1 + \frac{(f^0_{\theta^*} - f^0_\theta)^2}{H^0_\theta}\end{aligned}$$

Thus from (2.7.25),

$$\begin{aligned} 2\,A(m_0) &= \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}} - \log\Big(\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}}\Big) - 1 + \frac{\big(f^0_{\theta^*} - f^0_{\theta^*(m_0)}\big)^2}{H^0_{\theta^*(m_0)}}\Big] \\ &\geq \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}}\Big] - \log\Big(\mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}}\Big]\Big) - 1 + \mathbb{E}\Big[\frac{\big(f^0_{\theta^*} - f^0_{\theta^*(m_0)}\big)^2}{H^0_{\theta^*(m_0)}}\Big] \quad \text{by Jensen Inequality.} \end{aligned}$$

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

- If $f^0_{\theta^*} \neq f^0_{\theta^*(m_0)}$, then $\mathbb{E}\Big[\Big(\frac{f^0_{\theta^*} - f^0_{\theta^*(m_0)}}{M^0_{\theta^*(m_0)}}\Big)^2\Big] > 0$ and $A(m_0) > 0$.
- Otherwise, then

$$2\,A(m_0) = \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}} - \log\Big(\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}}\Big) - 1\Big],$$

  and from assumption $Id(\Theta)$, since $\theta^* \notin \Theta(m)$ and $f^0_{\theta^*} = f^0_{\theta^*(m_0)}$, we necessarily have $H^0_{\theta^*} \neq H^0_{\theta^*(m_0)}$ so that $\frac{H^0_{\theta^*}}{H^0_{\theta^*(m_0)}} \neq 1$. Therefore, $A(m_0) > 0$.

As a consequence,

$$\frac{\widehat{C}(m_0) - \widehat{C}(m^*)}{n} = A(m_0) + \frac{\kappa_n}{n}(|m_0| - |m^*|) + o_{a.s}(1). \tag{2.7.26}$$

Moreover, since $\kappa_n = o(n)$ and all the considered models are finite dimensional, the equality (2.7.26) 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 (2.2.4). 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 (2.1). 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 Bardet and Wintenberger (2009)), 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 2.1 which gives $\mathbb{P}\big(\widehat{m} \neq m^*\big) \underset{n\to\infty}{\longrightarrow} 0$.

□

2.7.3 Proof of Theorem 2.2

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

First, we have:

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

Under the assumptions of Theorem 2.1, $\mathbb{P}\big(\widehat{m} = m^*\big) \underset{n\to\infty}{\longrightarrow} 1$ and $\mathbb{P}\big(\widehat{m} \neq m^*\big) \underset{n\to\infty}{\longrightarrow} 0$. Therefore the second term in the right side of the previous equality asymptotically vanishes. For the first term, we can write,

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

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:

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

with Z a standard Gaussian random vector in $\mathbb{R}^{|m^*|}$ from the central limit theorem in Theorem 2 of Bardet and Wintenberger (2009), and this achieves the proof of 2.3.3 of Theorem 2.2. □

2.7.4 Proof of Theorem 2.3

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

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

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:

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

and $\rho_k(\theta) := \dfrac{\gamma_k(\theta)}{\gamma_0(\theta)}$, $\widehat{\rho}_k(\theta) := \dfrac{\widehat{\gamma}_k(\theta)}{\widehat{\gamma}_0(\theta)}$, $\rho_k(m) := \dfrac{\gamma_k(m)}{\gamma_0(m)}$ and $\widehat{\rho}_k(m) := \dfrac{\widehat{\gamma}_k(m)}{\widehat{\gamma}_0(m)}$.
Finally, for K a positive integer, denote the vector of adjusted correlogram:

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

Proof. (1) This proof is divided into two parts. In (i) we prove a result that ensures that the asymptotic distributions of the vectors $\widehat{\rho}(\theta)$ and $\rho(\theta)$ are the same. In (ii) we show that the large sample distribution of $\sqrt{n}\rho(m^*)$ is normal with a covariance matrix $V(\theta^*, m^*)$. Those two conditions do lead well to the asymptotic normality (2.5.1).

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

$$\sqrt{n}\left\|\widehat{\gamma}_k(\theta) - \gamma_k(\theta)\right\|_\Theta \xrightarrow[n\to\infty]{a.s.} 0. \tag{2.7.27}$$

We have:

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

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

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

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 (2.7.8) 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

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

Therefore,

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

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 Kounias and Weng (1969), $\|I_1^1\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0$ when

$$\sum_{t=1}^{\infty}t^{-1/4}\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)<\infty. \tag{2.7.28}$$

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

$$\sum_{t=1}^{\infty}t^{-1/4}\sum_{j\ge t}\alpha_j(\widetilde{H}_\theta,\Theta)<\infty. \tag{2.7.29}$$

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

$$\sum_{t=1}^{\infty}t^{-1/4}\Big(\sum_{j\ge t}\alpha_j(f_\theta,\Theta)+\alpha_j(M_\theta,\Theta)+\alpha_j(\widetilde{H}_\theta,\Theta)\Big)^{1/2}<\infty, \tag{2.7.30}$$

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 (2.7.22).
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 (2.7.30) that

$$\sqrt{n}\big\|\widehat{\rho}_k(\theta)-\rho_k(\theta)\big\|_\Theta \xrightarrow[n\to+\infty]{a.s.} 0 \quad \text{for any } k\ge 1. \tag{2.7.31}$$

This also implies

$$\sqrt{n}\big|\widehat{\rho}_k(m^*)-\rho_k(m^*)\big| \xrightarrow[n\to+\infty]{a.s.} 0 \quad \text{for any } k\ge 1. \tag{2.7.32}$$

(ii) The proof of this result has already been done in Li and Mak (1994) 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 $k = 1, \ldots, K$,

$$\sqrt{n}\,\gamma_k(m^*) = \sqrt{n}\,\gamma_k(\widehat{\theta}(m^*)) = \sqrt{n}\,\gamma_k(\theta^*) + \partial_\theta \gamma_k(\overline{\theta}^{(k)}) \sqrt{n}\,\big((\widehat{\theta}(m^*))_i - \theta_i^*\big)_{i\in m^*}, \quad (2.7.33)$$

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

$$\partial_\theta \gamma_k(\theta) = -\frac{2}{n}\Big(\sum_{t=k+1}^{n} e_t^2(\theta)\,\big(e_{t-k}^2(\theta) - 1\big)\frac{\partial_\theta M_\theta^t}{M_\theta^t} + e_t(\theta)\big(e_{t-k}^2(\theta) - 1\big)\,\frac{\partial_\theta f_\theta^t}{M_\theta^t} + e_{t-k}(\theta)\,\big(e_t^2(\theta) - 1\big)\frac{\partial_\theta f_\theta^{t-k}}{M_\theta^{t-k}} + e_{t-k}^2(\theta)\,\big(e_t^2(\theta) - 1\big)\frac{\partial_\theta M_\theta^{t-k}}{M_\theta^{t-k}}\Big). \quad (2.7.34)$$

We have $\mathbb{E}\big[e_{t-k}(\theta^*)\,\big(e_t^2(\theta^*) - 1\big)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}} \mid \sigma\big((\xi_s)_{s\le t-k}\big)\big] = e_{t-k}(\theta^*)\frac{\partial f_{\theta^*}^{t-k}}{M_{\theta^*}^{t-k}}\mathbb{E}\big[e_t^2(\theta^*) - 1\big] = 0$ since we have assumed $\mathbb{E}[\xi_0^2] = 1$. Moreover, $\mathbb{E}\big[e_t(\theta^*)\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big] = \mathbb{E}\big[\xi_t\,\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big] = 0$ and this implies $\mathbb{E}\big[e_t(\theta^*)\big(e_{t-k}^2(\theta^*) - 1\big)\,\frac{\partial f_{\theta^*}^t}{M_{\theta^*}^t}\big] = 0$. As a consequence, the expectation of the three last terms of (2.7.34) 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

$$\big(\partial_\theta \gamma_k(\overline{\theta}^{(k)})\big)_{1\le k\le K} \xrightarrow[n\to+\infty]{a.s.} J_K(m^*) = -2\,\Big(\mathbb{E}\Big[\big(\xi_0^2 - 1\big)\,\frac{\partial}{\partial\theta_j}\log\big(M_{\theta^*}^i\big)\Big]\Big)_{1\le i\le K,\, j\in m^*}. \quad (2.7.35)$$

Under the assumptions, a central limit theorem for $\widehat{\theta}(m^*)$ has been established in Bardet and Wintenberger (2009), and this implies

$$\big(\partial_\theta \gamma_k(\overline{\theta}^{(k)})\big)_{1\le k\le K}\sqrt{n}\,\big((\widehat{\theta}(m^*))_i - \theta_i^*\big)_{i\in m^*} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\Big(0\,,\; J_K(m^*)\,F(\theta^*,m^*)^{-1}G(\theta^*,m^*)F(\theta^*,m^*)^{-1}J_K'(m^*)\Big). \quad (2.7.36)$$

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:

$$\sqrt{n}\,\big(\gamma_k(\theta^*)\big)_{1\le k\le K} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\big(0\,,\,(\mu_4 - 1)^2\,I_K\big), \quad (2.7.37)$$

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

We would like to use (2.7.33) for obtaining the asymptotic behavior of $\gamma(m^*)$. In (2.7.36) and (2.7.37), 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 Bardet and Wintenberger (2009)), a Taylor expansion of log-likelihood for large n leads to

$$\big((\widehat{\theta}(m^*))_i - \theta_i^*\big)_{i\in m^*} \approx 2\,\big(F(\theta^*,m^*)\big)^{-1}\frac{1}{n}\,\frac{\partial}{\partial\theta}L_n(\theta^*).$$

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

$$-\,J_K(m^*)\,F(\theta^*,m^*)^{-1}\mathbb{E}\Big[\frac{\partial}{\partial\theta}L_n(\theta^*)\,\gamma(\theta^*)'\Big]. \tag{2.7.38}$$

From (2.2.1), 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,

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

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

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

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 $\mathbb{E}\Big[\xi_i\big(\xi_j^2-1\big)\big(\xi_{j-k}^2-1\big)\,\frac{\partial}{\partial\theta}f_{\theta^*}^i\Big]=0$ for any i,j and k. As a consequence,

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

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

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

Moreover the vector $\gamma(m^*)$ is normal distributed from Lemma 3.3 of Ling and Li (1997). 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 (2.5.1) holds with

$$V(\theta^*, m^*) := I_K + (\mu_4 - 1)^{-2}\, J_K(m^*)\, F(\theta^*, m^*)^{-1} G(\theta^*, m^*) F(\theta^*, m^*)^{-1} J'_K(m^*) \\ - 2\,(\mu_4 - 1)^{-1}\, J_K(m^*)\, F(\theta^*, m^*)^{-1}\, J'_K(m^*). \quad (2.7.39)$$

The proof is achieved after using the limit theorem (2.7.32).

(2) (2.5.2) follows directly from (2.5.1).

(3) We follow a same reasoning like in the proof of Theorem 2.2. For $x = (x_k)_{1\le k\le K} \in \mathbb{R}^K$, denote by $F_n(x) = \mathbb{P}\Big(\bigcap_{1\le k\le K} \sqrt{n}\left(\widehat{\rho}(\widehat{m})\right)_k \le x_k\Big)$ the distribution function of $\sqrt{n}\widehat{\rho}(\widehat{m})$. Applying the Total Probability Rule and by virtue of Theorem 2.1, we obtain:

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

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

3

General Hannan and Quinn Criterion for Common Time Series

The content of this chapter was taken from the article submitted for publication: K. KAMILA "*General Hannan and Quinn Criterion for Common Time Series* " https://arxiv.org/pdf/2101.04210.pdf .

Abstract

This chapter aims to study data driven model selection criteria for a large class of time

series, which includes ARMA or AR(∞) processes, as well as GARCH or ARCH(∞), APARCH and many others processes. We tackled the challenging issue of designing adaptive criteria which enjoys the strong consistency property. When the observations are generated from one of the aforementioned models, the new criteria, select the true model almost surely asymptotically. The proposed criteria are based on the minimization of a penalized contrast akin to the Hannan and Quinn's criterion and then involved a term which is known for most classical time series models and for more complex models, this term can be data driven calibrated. Monte-Carlo experiments and an illustrative example on the CAC 40 index are performed to highlight the obtained results.

3.1 Introduction

A common solution in model selection is to choose the model, minimizing a penalized based criterion which is the sum of two terms: the first one is the empirical risk (least squares, likelihood) that measures the goodness of fit and the second one is an increasing function of the complexity which aims to penalize large models and control the bias. Therefore a challenging task when designing a penalized criterion is the specification of the penalty term. Considering leading model selection criteria (BIC, AIC, Cp, HQ to name a few), one can see that the penalty term is a product of the model dimension with a sequence which is specific to the criteria. Indeed, a criterion is designed according to the goal one would like to achieve. The classical properties for model selection criteria include *consistency*, *efficiency* (oracle inequality, asymptotic optimality), *adaptative in the minimax sense*.

In this chapter, we focus on consistency property which aims at identifying the data generating process with high probability or almost surely. Hence, it requires the assumption whereby there exists a true model in the set of competitive models and the goal is to select this with probability approaches one as the sample size tends to infinity. In Bardet et al. (2020b), they studied model selection criteria regarding consistency in a large class of time series, which is the interest of this paper. The leading criterion obtained in this framework is the BIC; with a relatively heavy penalty, it ensures the selection of quite simple models. Moreover, several papers have established the consistency property in particular settings. For instance, Hannan and Quinn (1979) shows that the Hannan and Quinn (HQ) penalty $c \log\log n$ with $c > 2$ leads to a consistent choice of the true order in the framework of AR type models. One year later, Hannan (1980) (or Hannan and Deistler (2012)) extended this result for ARMA models.

Also, it has been proven in several contexts, that the BIC criterion Schwarz (1978) enjoys the consistency property: Shibata (1986) in the density estimation using hypothesis testing for autoregressive moving average models, Lebarbier and Mary-Huard (2004) in density estimation for independent observations, Bardet et al. (2020b) for a general class of time series, to name a few.

Compare to HQ penalty, the BIC penalty does not have the slowest rate of increase and then it can very often choose very simple models possible wrongs for small samples Hannan and Quinn (1979). Moreover, the HQ criterion has been derived for linear time series: AR models in Hannan and Quinn (1979), ARMA models in Hannan (1980) and Hannan and Deistler (2012). Is the HQ penalty still strongly consistent for heteroscedastic nonlinear models such as GARCH, APARCH or ARMA-GARCH? And what about a general class including linear and non linear models as well?

That raises a challenging question of designing robust penalties for most classical time

series models enjoying the model selection consistency. This is the issue we want to address in this chapter for a general class of times series called affine causal and defined below.

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

$$X_t = M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t + f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \text{ for any } t\in\mathbb{Z}. \tag{3.1.1}$$

where $(\xi_t)_{t\in T}$ is a sequence of zero-mean independent identically distributed random vectors (i.i.d.r.v) satisfying $\mathbb{E}(|\xi_0|^r)<\infty$ with $r\geq 1$ and M, $f:\mathbb{R}^\infty\to\mathbb{R}$ are two measurable functions.

For instance,

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

Note that, numerous classical time series models such as ARMA(p,q), GARCH(p,q), ARMA(p,q)-GARCH(p,q) (see Ding et al. (1993) and Ling and McAleer (2003)) or APARCH(δ,p,q) processes (see Ding et al. (1993)) belongs to $\mathcal{AC}(M,f)$.
The study of this type of process more often requires the classical regularity conditions on the functions M and f, which are not restrictive at all and remain valid in various time serie models. Let us recall these conditions for $\Psi_\theta=f_\theta$ or M_θ and Θ a compact set.

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

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

In addition, if the noise ξ_0 admits r-order moments (for $r\geq 1$), let us define:

$$\Theta(r)=\Big\{\theta\in\mathbb{R}^d,\ A(f_\theta,\{\theta\}) \text{ and } A(M_\theta,\{\theta\}) \text{ hold with } \sum_{k=1}^\infty\alpha_k(f_\theta,\{\theta\})+\|\xi_0\|_r\sum_{k=1}^\infty\alpha_k(M_\theta,\{\theta\})<1\Big\}. \tag{3.1.2}$$

Under this assumption, Doukhan and Wintenberger (2008) showed that there exists a stationary and ergodic solution to (3.1.1) with r-order moment for any $\theta\in\Theta(r)$. Moreover, Bardet and Wintenberger (2009) studied the consistency and the asymptotic normality of the QMLE of θ^* for $\mathcal{AC}(M_{\theta^*},f_{\theta^*})$.

The main contribution of this chapter is the generalization of the HQ criterion to affine causal class: we provide a minimal multiplicative penalty term c_{min} so that all penalties of the form $2\,c\,\log\log n\,D_m$ with $c\geq c_{min}$ ensure the strong consistency property for affine causal models under some mild conditions on the Lipschitz coefficients of functions M_θ, f_θ (D_m denotes the size of the model m). Monte Carlo experiments have been conducted in order to attest the accuracy of our new criteria.

The chapter is organized as follows. The model selection consistency along with notations and assumptions are described in Section 3.2. Numerical results are presented in Section 3.3 and Section 3.4 contains the proofs.

3.2 Model Selection Consistency

3.2.1 Model Selection Procedure

Let assume $(X_1, \dots, X_n)$ be a trajectory of a stationary affine causal process $m^* := \mathcal{AC}(M_{\theta^*}, f_{\theta^*})$, where θ^* is unknown. The goal of the consistency property is to come up with this true model given a set of candidate model $\mathcal{M}$ such that $m^* \in \mathcal{M}$.

A D_m-dimensionnal model $m \in \mathcal{M}$ can be viewed as a set of causal functions (M_θ, f_θ) with $\theta \in \Theta(m) \subset \mathbb{R}^{D_m}$. $\Theta(m)$ is the parameter set of the model m.
The consistency property will be studied using quasi likelihood functions since assumption on the distribution of the noise is not required.
The Gaussian quasi log-likelihood is derived from the conditional (with respect to the filtration $\sigma\big(X_t, t \leq 0\big)$ log-likelihood of $(X_1, \dots, X_n)$ when (ξ_t) is supposed to be a Gaussian standard white noise. From (3.1.1), one deduce that the log density of X_t given $\sigma\big(X_i, i < t\big)$ is

$$-\frac{1}{2}\Big[\frac{(X_t - f^t_{\theta^*})^2}{H^t_{\theta^*}} + \log(H^t_{\theta^*})\Big].$$

Therefore the conditional log density of $(X_1, \dots, X_n)$ given $\sigma\big(X_t, t \leq 0\big)$ is

$$-\frac{1}{2}\sum_{t=1}^{n}\Big[\frac{(X_t - f^t_{\theta^*})^2}{H^t_{\theta^*}} + \log(H^t_{\theta^*})\Big].$$

From now on, we drop the Gaussian assumption of the noise. The conditional log-density inspires to define for all $\theta \in \Theta$

$$L_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} q_t(\theta)\ , \text{ with } q_t(\theta) := \frac{(X_t - f^t_\theta)^2}{H^t_\theta} + \log(H^t_\theta) \tag{3.2.1}$$

where $f^t_\theta := f_\theta(X_{t-1}, X_{t-2}, \cdots)$, $M^t_\theta := M_\theta(X_{t-1}, X_{t-2}, \cdots)$ and $H^t_\theta = \big(M^t_\theta\big)^2$. The quasi likelihood function L_n is not computable since it depends on the past $(X_{-j})_{j\in\mathbb{N}}$ that is unknown. However, the sequence $(L_n(.))_n$ enjoys very nice asymptotic properties such that the Uniform Law of Large Numbers (see Bardet and Wintenberger (2009)).
Let $\mathcal{M}$ a finite family of candidate models containing the true one m^*. According to Proposition 1 in Bardet et al. (2020b), all these models can be included into a big one with parameter space Θ. For each specific model $m \in \mathcal{M}$, we define the Gaussian QMLE $\widehat{\theta}(m)$ with respect to L_n as

$$\widehat{\theta}(m) = \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\ L_n(\theta). \tag{3.2.2}$$

To select the true model $m \in \mathcal{M}$, we consider a penalized contrast $C(m)$ ensuring a trade-off between -2 times the maximized 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

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ C(m) \quad \text{with} \quad C(m) = -2\,L_n\big(\widehat{\theta}(m)\big) + \kappa_n(m) \tag{3.2.3}$$

where $(\kappa_n)_n$ an increasing sequence depending on the number of observations n and the dimension D_m. There exist several possible choices of $\kappa_n(m)$ including

- $\kappa_n(m) = 2c\,D_m\,\log\log n$ with $c > 1$, we retrieve the HQ criterion Hannan and Quinn (1979);

- $\kappa_n = D_m \log n$, C yields to BIC criterion Schwarz (1978);
- $\kappa_n = 2\, D_m$, C is the AIC criterion Akaike (1973).

Basically the principle is that by increasing the size of the model, the likelihood increases also. The question is whether this increase in complexity is offset by a sufficient increase in likelihood. If the answer is no, then the least complex model is used, even if it is less likely. If the answer is affirmative, then we accept to work with a more complex model. Of course, all the difficulty lies in the choice of weights between likelihood and complexity, and thus ultimately in the specification of the penalty multiplicative term κ_n.

What is the better weighting term of the model complexity? The aim here is by leveraging the increasing rate of the likelihood, to propose a data driven κ_n in order to guarantee the strong consistency property to our model selection procedure i.e.

$$\widehat{m} \xrightarrow[n\to\infty]{a.s.} m^*. \tag{3.2.4}$$

3.2.2 Assumptions

Some mild conditions will be required to prove the consistency of the considered model selection criteria.

The following assumption is well-known as the identifiability one and is always required in order to guarantee the unicity of the global maximum of the MLE at the true parameter θ^*. That is:

Assumption A1: *For all* θ, $\theta' \in \Theta_m$, $(f_\theta^0 = f_{\theta'}^0)$ *and* $(M_\theta^0 = M_{\theta'}^0) \implies \theta = \theta'$.

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 Bardet and Wintenberger (2009), Straumann and Mikosch (2006), White (1982)).

The following condition provides the invertibility of the Fisher's information matrix of $(X_1, \ldots, X_n)$ and was used to prove the asymptotic normality of the QMLE (see Bardet and Wintenberger (2009)).

Assumption A2: *For any* $x \in \mathbb{R}^\infty$, *the functions* $\theta \to M_\theta$ *and* $\theta \to f_\theta$ *are* $\mathcal{C}^2(\Theta)$ *and one of the families* $(\partial f_\theta^t / \partial \theta^{(i)})_{1\le i\le D_{m^*}}$ *or* $(\partial H_\theta^t / \partial \theta^{(i)})_{1\le i\le D_{m^*}}$ *is a.e. linearly independent.*

Note that the definitions of the conditional log-likelihood requires 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 A3: $\exists \underline{h} > 0$ *such that* $\inf_{\theta\in\Theta}(H_\theta(x)) \ge \underline{h}$ *for all* $x \in \mathbb{R}^\infty$.

Next we assume the existence of the eighth order moment of the noise.

Assumption A4: $\mathbb{E}[\xi_0^8] < \infty$.

We end the list of assumptions by assuming a suitable relation between the Fisher Information matrix $G(\theta_m^*)$ and the limiting Hessian matrix of the log-likelihood $F(\theta_m^*)$ defined as follows

$$\big(F(\theta_m^*)\big)_{i,j} = \mathbb{E}\Big[\frac{\partial^2 q_0(\theta_m^*)}{\partial\theta_i \partial\theta_j}\Big] \quad \text{and} \quad (G(\theta_m^*))_{i,j} = \mathbb{E}\Big[\frac{\partial q_0(\theta_m^*)}{\partial\theta_i}\frac{\partial q_0(\theta_m^*)}{\partial\theta_j}\Big],$$

with $\theta_m^* := (\theta^*, 0, \dots, 0)^\top \in \Theta(m)$.

Assumption A5: There exist absolutes constants α_1 and α_2 such that for any $m \in \mathcal{M}$ verifying $m^* \subset m$,

$$\mathbf{1}_m^\top \Sigma_{\theta_m^*} \mathbf{1}_m = \alpha_1 D_m^1 + \alpha_2 D_m^2 \tag{3.2.5}$$

where D_m^1 and D_m^2 are two integers such that $D_m^1 + D_m^2 = D_m$, $\quad \mathbf{1}_m := (1, 1, \dots, 1)^\top \in \mathbb{R}^{D_m}$, $\Sigma_{\theta_m^*} := G(\theta_m^*)^{1/2} F(\theta_m^*)^{-1} G(\theta_m^*)^{1/2}$.

For most classical affine causal models, **A5** is verified (see Proposition 3.2). However, for more complex models such as ARMA-GARCH with $\mu_4 \neq 3$, $\Sigma_{\theta_m^*}$ is hard to handle.

3.2.3 Consistency Result

Before stating the main result of this section, we give important intermediate results. All proof of the results stated in this subsection can be found in Section 3.4.

The following Proposition suggests the existence of a term that will be the keystone of this work.

Proposition 3.1. *Let m^* any affine causal model. For any model m verifying $m^* \subset m$, and under **A1-A5**, it holds*

$$\limsup_{n\to\infty} \frac{L_n\big(\widehat{\theta}(m)\big) - L_n(\theta_m^*)}{2 \log\log n} = \frac{1}{4}\big(\alpha_1 D_m^1 + \alpha_2 D_m^2\big) \quad a.s. \tag{3.2.6}$$

*where the constants α_1, α_2, D_m^1, D_m^2 are specified in assumption **A5**.*

For every $m \in \mathcal{M}$, let us denote by $c_{min}(m)$ the following term that will be used several times

$$c_{min}(m) := \frac{1}{4}\big(\alpha_1 D_m^1 + \alpha_2 D_m^2\big) \tag{3.2.7}$$

Now we state a result which provides the values of both α_1 and α_2 for most classical affine causal models.

Proposition 3.2. *Under the assumptions and notation of Proposition 3.1, we have*

- *If $\mu_4 = \mathbb{E}[\xi_0^4] = 3$ (for instance for Gaussian noise), then $\alpha_1 = 2$, $\alpha_2 = 2$ and $c_{min}(m) = \frac{1}{2} D_m$;*
- *If the parameter θ identifying an affine causal model $X_t = M_\theta^t \xi_t + f_\theta^t$ can be decomposed as $\theta = (\theta_1, \theta_2)'$ with $f_\theta^t = \widetilde{f}_{\theta_1}^t$ and $M_\theta^t = \widetilde{M}_{\theta_2}^t$, then $\alpha_1 = 2$, $\alpha_2 = \mu_4 - 1$ and*

$$c_{min}(m) = \frac{1}{2} D_m^1 + \frac{\mu_4 - 1}{4} D_m^2$$

The second configuration in Proposition 3.2 includes classical time series

- GARCH(p, q), APARCH(δ, p, q) type models and related ones, $c_{min}(m) = \frac{\mu_4 - 1}{4} D_m$;
- ARMA(p, q) models, $c_{min}(m) = \frac{D_m}{2}$ if the variance of the noise is known and $c_{min}(m) = \frac{D_m - 1}{2} + \frac{\mu_4 - 1}{4}$ otherwise.

We can now state the first main result of this paper.

Theorem 3.1. *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 belonging to $\Theta(r) \subset \mathbb{R}^{D_{m^*}}$. Let also $\mathcal{M}$ be a finite family of candidate models such that $m^* \in \mathcal{M}$. If assumptions **A1-A5** hold, then with $c_{min} := \max\left(\frac{\alpha_1}{2}, \frac{\alpha_2}{2}\right)$, it holds*

for any $\kappa_n(m) = 2\,c\,D_m \log\log n$ with

$$c \geq c_{min} \tag{3.2.8}$$

it holds for the selected model $\widehat{m}$ according to (3.2.3)

$$\widehat{m} \underset{n\to\infty}{\xrightarrow{a.s.}} m^*, \tag{3.2.9}$$

Remark 3.1.

1. *For classical configurations as seen in Proposition 3.2, this result gives a generalization of Hannan and Quinn criterion.*
2. *For more complex models, the values of α_1 and α_2 are unknowns (at least until a better relationship between matrix $F(\theta_m^*)$ and $G(\theta_m^*)$ is found) and so c_{min} is also unknown. In these cases, we propose to use adaptive methods such as slope heuristic algorithm or dimension jump Arlot and Massart (2009) to calibrate c_{min}.*

Let us mention that our result generalizes the strong consistency obtained by Hannan and Quinn (1979) for AR models as the affine causal class also contains GARCH models. It furthermore generalizes the result Hannan and Deistler (2012) for ARMA models.

Theorem 3.1 gives a theoretical guarantee on the consistency of the model selection procedure. However, it does not say anything about the convergence (and its rate) of the parameter estimate resulting from the model selection $\widehat{m}$. The following results shows that the final estimate $\widehat{\theta}_{\widehat{m}}$ is consistent and verifies a CLT.

Theorem 3.2. *Under the assumptions of Theorem 3.1, it holds*

$$\widehat{\theta}(\widehat{m}) \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} \theta^*. \tag{3.2.10}$$

Moreover,

$$\sqrt{n}\left(\left(\widehat{\theta}(\widehat{m})\right)_i - (\theta^*)_i\right)_{i\in m^*} \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}_{|m^*|}\left(0\ , \Sigma_{\theta^*,m^*}\right) \tag{3.2.11}$$

with $\Sigma_{\theta^*,m^*} := F(\theta^*)^{-1}G(\theta^*)F(\theta^*)^{-1}$

The proof of this result is identical to the proofs of theorems 3.1 and 3.2 of our previous paper Bardet et al. (2020b). It will therefore not be carried out for the sake of brevity.

In this subsection, we have used the QMLE contrast without any distribution assumption on the noise to derive a consistency property. However, the contrast L_n as in (3.2.1) depends on all the past values of the process X, which are unobserved. In the next subsection, we will propose an extension of Theorem 3.1 based on QMLE which does not require knowledge of the initial values of the process.

3.2.4 Another Quasi-Likelihood Function

The goal of this subsection is to sharpen the conditions on the sequence κ_n found in Bardet et al. (2020b). Before stating the result, let recall a little bit some definitions and notations used in Bardet et al. (2020b). Following the derivation of L_n, we define its computable version $\widehat{L}_n$ as follows:

$$\widehat{L}_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} \widehat{q}_t(\theta) \ , \ \text{with } \widehat{q}_t(\theta) := \frac{(X_t - \widehat{f}_\theta^t)^2}{\widehat{H}_\theta^t} + \log(\widehat{H}_\theta^t) \tag{3.2.12}$$

where $\widehat{f}_\theta^t := f_\theta(X_{t-1}, X_{t-2}, \ldots, X_1, 0, \ldots, 0)$, $\widehat{M}_\theta^t := M_\theta(X_{t-1}, X_{t-2}, \ldots, X_1, 0, \ldots, 0)$ and $\widehat{H}_\theta^t = \big(\widehat{M}_\theta^t\big)^2$.
Therefore for every model $m \in \mathcal{M}$, we define the Gaussian QMLE $\widetilde{\theta}(m)$ (with respect to $\widehat{L}_n$) as

$$\widetilde{\theta}(m) = \underset{\theta\in\Theta(m)}{\operatorname{argmax}}\ \widehat{L}_n(\theta). \tag{3.2.13}$$

Once the estimation in each model in the family $\mathcal{M}$ has been performed, we select the best model as follows

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \widehat{C}(m) \quad \text{with} \quad \widehat{C}(m) = -2\,\widehat{L}_n\big(\widetilde{\theta}(m)\big) + \kappa_n(m). \tag{3.2.14}$$

In this framework, we do not consider long memory process and then we define the class $\mathcal{H}(M_\theta, f_\theta)$ a subset of $\mathcal{AC}(M_\theta, f_\theta)$ in which every process has Lispchitz coefficients satisfying the following conditions

$$\alpha_j(f_\theta) + \alpha_j(M_\theta) + \alpha_j(\partial_\theta f_\theta) + \alpha_j(\partial_\theta M_\theta) = O(j^{-\gamma}) \quad \text{with} \quad \gamma > 2.$$

It is then straightforward to see that every process in the class $\mathcal{H}(M_\theta, f_\theta)$ verifies the following condition

Condition $\boldsymbol{K}(\Theta)$:

$$\sum_{k\geq e} \frac{1}{\log\log k} \sum_{j\geq k} \alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta) < \infty.$$

This finding allows to propose a sharpen generalization of both Theorem 3.1 in our previous paper Bardet et al. (2020b) and a similar result in Kengne (2020).

Before stating the consistency result, it is important as in Bardet et al. (2020b) or Bardet and Wintenberger (2009), to distinguish the special case of NLARCH(∞) processes which includes for instance GARCH(p, q) or ARCH(∞) processes. In such case, let us define the class:

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

$$X_t = \xi_t \sqrt{\widetilde{H}_\theta\big((X_{t-i}^2)_{i\in\mathbb{N}^*}\big)} \ \text{ for any } t \in \mathbb{Z}. \tag{3.2.15}$$

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

$$\widetilde{\Theta}(r) = \Big\{\theta \in \mathbb{R}^d,\ A(\widetilde{H}_\theta, \{\theta\}) \text{ holds with } \big(\|\xi_0\|_r\big)^2 \sum_{k=1}^{\infty} \alpha_k(\widetilde{H}_\theta, \{\theta\}) < 1\Big\}. \tag{3.2.16}$$

Finally, we propose to restrict class $\widetilde{\mathcal{AC}}(\widetilde{H}_\theta)$ to $\widetilde{\mathcal{H}}(\widetilde{H}_\theta)$ as done with $\mathcal{H}(M_\theta, f_\theta)$ by considering all the process checking the condition

$$\alpha_j(H_\theta) + \alpha_j(\partial_\theta H_\theta) = O(j^{-\gamma}) \quad \text{with} \quad \gamma > 2$$

so that

$$\sum_{k\geq e} \frac{1}{\log\log k} \sum_{j\geq k} \alpha_j(H_\theta) + \alpha_j(\partial_\theta H_\theta) < \infty.$$

We can now state the second main result.

Theorem 3.3. *Let $(X_1, \ldots, X_n)$ be an observed trajectory of an affine causal process X belonging to $\mathcal{H}(M_{\theta^*}, f_{\theta^*})$ (or $\widetilde{\mathcal{H}}(\widetilde{H}_{\theta^*})$) where θ^* is an unknown vector belonging to $\Theta(r) \subset \mathbb{R}^{D_{m^*}}$ (or $\widetilde{\Theta}(r) \subset \mathbb{R}^{D_{m^*}}$). Let also $\mathcal{M}$ be a finite family of candidate models such that $m^* \in \mathcal{M}$. If assumptions **A1-A5** hold, then with $c_{min} := \max\left(\frac{\alpha_1}{2}, \frac{\alpha_2}{2}\right)$, it holds*

for any $\kappa_n(m) = 2\,c\,D_m\,\log\log n$ with

$$c \geq c_{min} \tag{3.2.17}$$

it holds for the selected model $\widehat{m}$ according to (3.2.14)

$$\widehat{m} \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} m^*. \tag{3.2.18}$$

All the comments made about the Theorem 3.1 remain valid here. Moreover, recently, Kengne (2020) requires heavy penalties to ensure the strong consistency for the process in the class $\mathcal{AC}(M_{\theta^*}, f_{\theta^*})$. Indeed, according to Kengne (2020), it is necessary that κ_n verified $\kappa_n/\log\log n \underset{n\to\infty}{\longrightarrow} \infty$ to obtain (3.2.18) which is a stronger condition since the HQ criterion does not fulfill this condition and it is well known that HQ criterion is strongly consistent (see for instance Hannan and Quinn (1979)). Moreover, the new penalties found in this paper does not satisfy this condition, yet we ensure strongly consistency.

Also, let us mention that for small samples, heavy penalties such as those in Kengne (2020) can very often choose very simple models possible wrongs Hannan and Quinn (1979).

Remark 3.2. *Our results show that, asymptotically heavy penalties such as BIC penalty will ensure the consistency property. However, in practise, these penalties are used for fixed n and most for small samples and it is important to point out their drawbacks. Very often, in the family of competitive models $\mathcal{M}$, there are misspecified and underfitted models ($\mathcal{M}'$). Since the difference $-2\,\widehat{L}_n(\widehat{\theta}_m) + 2\,\widehat{L}_n(\widehat{\theta}_{m^*}) > 0$ for every $m \in \mathcal{M}'$, making the penalty, heavy could offset this positivity and can lead to the selection of some underfitted models and then wrong models. To be more convincing of that, see the simulations (DGP III) experiments in Section 3.3.*

3.2.5 Algorithm of Calibration of the minimal constant

There exist several ways to calibrate the minimal constant c_{min} including the dimension jump (presented below) and the data-driven slope estimation. Indeed, once an estimation of c_{min} is obtained, many studies advocates the choice of $2\,\widehat{c}_{min}$ which turn out to be optimal (Massart (2007), Arlot and Massart (2009) among others). Now we present the dimension jump algorithm.

Algorithm 3.1. ***Dimension Jump*** ***Arlot and Massart (2009)***

1. *Compute the selected model* $\widehat{m}(c)$ *as a function of* $c > 0$

$$\widehat{m}(c) = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \, \widehat{L}_n\big(\widehat{\theta}_m\big) + c \, \mathrm{pen}_{shape}(m); \quad \mathrm{pen}_{shape}(m) = D_m \, \log\log n$$

2. *Find* $\widehat{c}_{min}$ *such that* $D_{\widehat{m}(c)}$ *is "huge" for* $c < \widehat{c}_{min}$ *and "reasonnably small" for* $c \geq \widehat{c}_{min}$*;*

3. *Select the model* $\widehat{m} := \widehat{m}(2\,\widehat{c}_{min})$.

This algorithm has been implemented in Baudry et al. (2012) which gives several details including the grid for c values.

Let us notice that, in view of obtaining penalties, there is no need to calibrate the c_{min} constant for most classical time series models. However, since the fourth order moment of the noise is unknown, a consistent estimate of this term is required. To do that, we proceed as in the estimation of the variance of the noise as in the Mallows Cp.

A consistent estimator $\widehat{\mu}_{\overline{m},4}$ of $\mu_4 = \mathbb{E}[\xi_0^4]$ can be :

$$\widehat{\mu}_{\overline{m},4} := \frac{1}{n} \sum_{t=1}^{n} \big(\widehat{\xi}_t(\overline{m})\big)^4 \quad \text{with} \quad \widehat{\xi}_t(\overline{m}) := \big(\widehat{M}^t_{\widehat{\theta}_{\overline{m}}}\big)^{-1} \big(X_t - \widehat{f}^t_{\widehat{\theta}_{\overline{m}}}\big)$$

where we suppose that $\overline{m}$ is the "largest" model in the family $\mathcal{M}$, typically the largest order of a family of time series. As a result an estimator of the c_{min} constant to consider in the penalty κ_n is

- $\frac{\widehat{\mu}_{\overline{m},4} - 1}{2}$ for GARCH family and related ones;
- 1 for ARMA family with known variance.

3.3 Numerical Experiments

In this section, several numerical experiments are conducted to assess the consistency property (Section 3.2) of our new criteria.

3.3.1 Monte Carlo: Consistency

This subsection studies the performance of the model selection criteria found in Section 3.2. We have considered three different Data Generating Process (DGP):

$$\begin{aligned}
\text{DGP I} \quad & X_t = 0.5\, X_{t-1} + 0.2\, X_{t-2} + \xi_t, \\
\text{DGP II} \quad & X_t = \big(0.2 + 0.4\, X_{t-1}^2 + 0.2\, X_{t-2}^2\big)^{1/2} \xi_t, \\
\text{DGP III} \quad & X_t = 0.1(\, X_{t-1} + X_{t-2} + \ldots + X_{t-6}) + \xi_t,
\end{aligned}$$

where (ξ_t) will be a white Gaussian noise with variance one at first and a Student with 5 degrees of freedom on the other hand. For the first and the second DGP, we considered as competitive models all the models in the family $\mathcal{M}$ defined by:

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

Therefore, there are 66 candidate models as in Bardet et al. (2020b). The goal is to compare the ability of selecting the true model for $\kappa_n^3 = \log n \, D_m$, $\kappa_n^2 = 2\,\widehat{c}_{min} \, \log\log n \, D_m$

(in accordance with the condition (1.3.12) and $\kappa_n^2 = \mathbf{2} \times 2\,\widehat{c}_{min}\,\log\log n\,D_m$. Moreover, from Proposition 1.4, $\widehat{c}_{min}$ does not need to be estimated and worth one for Gaussian noise. But for Student noise, $\widehat{c}_{min} = \max\left(1, \frac{\widehat{\mu}_{\overline{m},4}-1}{2}\right)$ for the DGP I and $\widehat{c}_{min} = \frac{\widehat{\mu}_{\overline{m},4}-1}{2}$ for DGP II. The Table 2.1 presents the results of the selection procedure. As we can notice, the three penalties have a good consistency property. Moreover, for n relatively small, the penalty κ_n^1 is better than both others. For larger n, κ_n^2 is the best the penalty to consider.

Table 3.1: Percentage of selected order based on 500 independent replications depending on sample's length for the penalty $\kappa_n^1 = 2\,\widehat{c}_{min}\,\log\log n\,D_m$, $\kappa_n^2 = 4\,\widehat{c}_{min}\,\log\log n\,D_m$ and $\kappa_n^3 = \log n\,D_m$, where and W, T, O refers to wrong, true and overfitted selection.

	n	100			500			1000			2000		
		κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3
	W	58.2	80.8	94.6	23.6	26	36.2	18.8	17.6	17.8	4.2	7.4	7.8
DGP I	T	32.6	19	5.4	64.2	73.4	63.8	71.6	82.2	82.0	89.4	92.2	92.2
Gaussian	O	9.2	0.2	0	12.2	0.6	0	9.6	0.2	0.2	6.4	0.5	0
	W	70.0	84.8	94.0	17.4	21.6	35.4	16.4	15.8	15.8	3.6	3.6	5
DGP II	T	29.8	15.2	6.0	81.0	78.4	64.6	83.4	84.2	84.2	96.4	96.4	95
Gaussian	O	0.2	0	0	1.6	0	0.	0.2	0	0.0	0	0	0
	W	70.1	86.8	97.5	38.3	53.1	35.0	16.0	19.5	16.3	18	22.0	18
DGP I	T	19.8	13.2	2.5	59.7	46.9	65.9	83.9	80.5	82.7	81.4	78.0	82.0
Student	O	10.1	0	0.	2.0	0	0.1	0.1	0	0	0.6	0	0
	W	75.0	84.2	88.8	45.6	57.2	55.4	25.6	32.6	26.2	14	16.0	13.0
DGP II	T	22.4	15.0	11.2	49.0	41.2	44.6	71.0	66.8	73.8	85.0	84.0	87.0
Student	O	2.6	0.8	0.0	4.4	1.6	1.0	3.4	0.6	0.0	1.0	0	0

For DGP III, as we want to exhibit the possible "non consistency" of BIC for small samples, we have considered as the competitive set, the hierarchical family of AR models

$$\mathcal{M}' = \{AR(1), \ldots, AR(15)\}$$

for $n = 100, 200, 400, 500, 1000, 2000$. In Table 3.2 below the percentage of selected order based on 1000 independent replications are presented for the above three penalties.

Table 3.2: Percentage of selected order based on 1000 independent replications depending on sample's length using penalty terms $\kappa_n^1 = \widehat{c}_{min}$, $\kappa_n^2 = 2\,\widehat{c}_{min}$ and $\kappa_n^3 = \log n$, for DGP III.

n	100			200			400			500			1000			2000		
	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3	κ_n^1	κ_n^2	κ_n^3
$p < 6$	61.5	96	99	46	83.5	97	29.5	59	87.5	19.5	45.5	76.5	3.0	9.5	38.5	0	0.5	5.0
$p = 6$	15	3.0	1.0	25	13	2.5	44	37.5	12.5	50.5	49.5	22.5	68	86.5	61	72.5	96	94.5
$p > 6$	23.5	1.0	0	29	3.5	0.5	26.5	3.5	0	30	5.0	1.0	29	4.0	0.5	27.5	3.5	0.5

These results invite us to be cautious when using the BIC for small sample sizes, whereas the proposed adaptive penalty is more robust, as it at least allow us to recover an overfitted model that is less harmful than a wrong model most often chosen by the BIC.

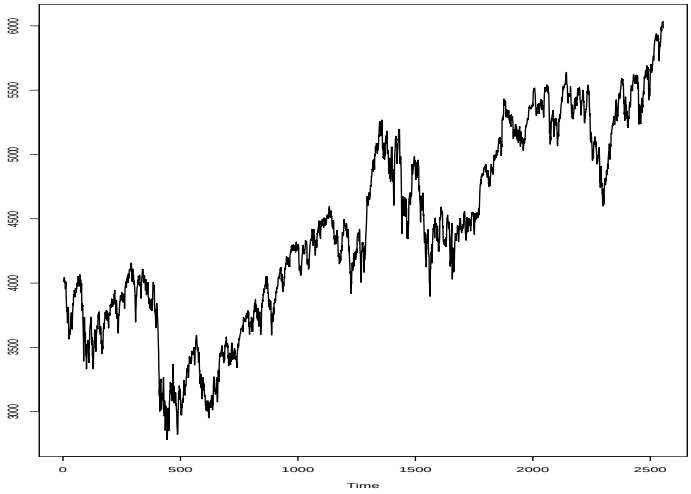


Figure 3.1: Daily closing CAC 40 index (January 1st, 2010 to December 31 st, 2019).

3.3.2 Real Data Analysis: financial time series

CAC 40 is a benchmark French stock market index and is highly regarded in many statistical studies . Let consider the daily closing prices of the CAC 40 index from January 1st, 2010 to December 31st, 2019 plotted in Figure 3.1. Over the period under review, the CAC 40 increased.

To analyze this type of data, it is common to consider the returns (see Figure 3.2). We can see that the return values display some small auto-correlations. Also, from Figure 3.3, the squared returns of CAC 40 are strongly auto-correlated. These facts suggest that the strong white noise assumption cannot be sustained for this log-returns sequence of the CAC 40 index.

Hence, let consider the competitive set of models $\mathcal{M}$ used in the previous subsection in order to propose the best suitable model for these data:

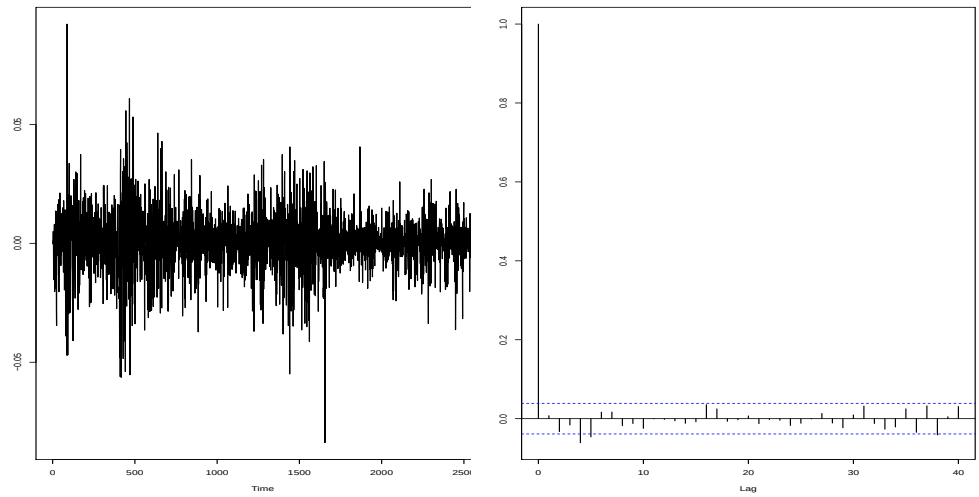
$$\mathcal{M} = \big\{\text{ARMA}(p,q) \text{ or GARCH}(p',q') \text{ processes with } 0 \leq p,q,p' \leq 5,\ 1 \leq q' \leq 5\big\}.$$

Using the adaptive penalty and the BIC criterion, we find out that the GARCH$(1,1)$ is the best model over $\mathcal{M}$ with respect to both criteria. This fact is in accordance with Francq and Zakoian (2010) which found the same result using the returns of the CAC 40 index from March 2, 1990 to December 29, 2006.

3.4 Proofs

3.4.1 Proof of Theorem 3.1

Usually, this proof is divided into two parts: one has to show that as n tends to infinity the probability of overfitting goes to zero and so is the probability of misspecification.



(a) Time plot of log returns. **(b)** Correlograms of log returns.

Figure 3.2: Daily closing CAC 40 index (January 4th, 2010 to December 31 st, 2019).

3.4.1.a Overfitting Case

Let $m \in \mathcal{M}$ such as $m^* \subset m$. We want to show that $C(m^*) \leq C(m)$ a.s. asymptotically. We have

$$\begin{aligned} C(m^*) \leq C(m) &\iff -2\,L_n\big(\widehat{\theta}(m^*)\big) + 2c\,\log\log n\,D_{m^*} \leq -2\,L_n\big(\widehat{\theta}(m)\big) + 2c\,\log\log n\,D_m \\ &\iff \frac{L_n\big(\widehat{\theta}(m)\big) - L_n(\theta^*)}{\log\log n} \leq \frac{L_n\big(\widehat{\theta}(m^*)\big) - L_n(\theta^*)}{\log\log n} + \quad c\,(D_m - D_{m^*}) \end{aligned} \tag{3.4.1}$$

therefore, a necessary and sufficient condition to avoid overfitting can be stated by taking $\limsup_{n\to\infty}$ on both sides of the inequality (3.4.1); that is by virtue of definition (1.3.11)

$$2\,c_{min}(m) - 2\,c_{min}(m^*) \leq c\,(D_m - D_{m^*}) \quad \text{for} \quad m^* \subset m, \tag{3.4.2}$$

i.e.,

$$\frac{\alpha_1}{2}\,(D^1_m - D^1_{m^*}) + \frac{\alpha_2}{2}\,(D^2_m - D^2_{m^*}) \leq c\,(D_m - D_{m^*})$$

which is fulfilled for any constant c such as in (3.2.8). Indeed, $c_{min} = \max\left(\frac{\alpha_1}{2}, \frac{\alpha_2}{2}\right)$ and

$$\begin{aligned} \frac{\alpha_1}{2}\,(D^1_m - D^1_{m^*}) + \frac{\alpha_2}{2}\,(D^2_m - D^2_{m^*}) &\leq c_{min}\Big(D^1_m - D^1_{m^*} + D^2_m - D^2_{m^*}\Big) \\ &= c_{min}(D_m - D_{m^*}) \end{aligned}$$

where the inequality holds since $m^* \subset m$ that implies $D^1_m \geq D^1_{m^*}$ and $D^2_m \geq D^2_{m^*}$. Hence the associated criterion κ_n will avoid overfitting.

3.4.1.b Misspecification/Underfitting Case

All misspecified/underfitted models are contained in the set

$$\mathcal{M}' = \Big\{m \in \mathcal{M} : (m^* \not\subseteq m) \cup (m \subset m^*)\Big\}.$$

The proof is exactly as the one done in Bardet et al. (2020b). But for the sake of completeness, we give here some important steps of the proof.

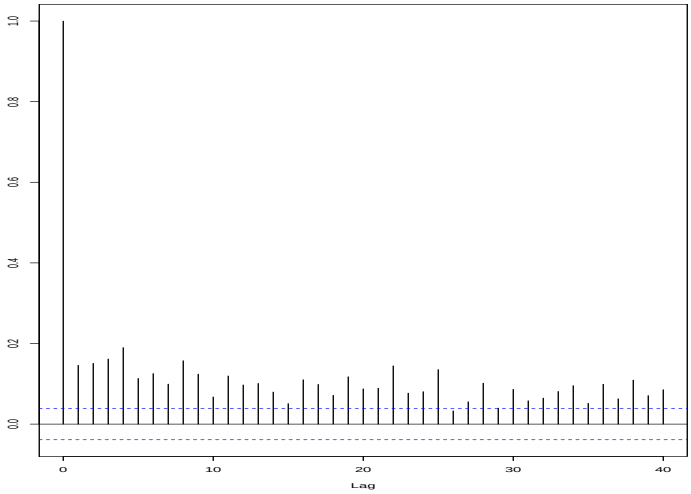


Figure 3.3: Sample autocorrelations of squared returns of the CAC 40 index (January 1st, 2010 to December 31 st, 2019).

The goal is to show that for every $m \in \mathcal{M}'$, it holds

$$C(m^*) \quad < \quad C(m) \quad a.s. \tag{3.4.3}$$

Let $m \in \mathcal{M}'$. From Proposition 2 in Bardet et al. (2020b) and using continuous mapping Theorem

$$\frac{1}{n}\Big[L_n\big(\widehat{\theta}(m^*)\big) - L_n\big(\widehat{\theta}(m)\big)\Big] = A(m_0) + o_{a.s}(1) \tag{3.4.4}$$

where

$$A(m) := L(\theta^*) - L(\theta^*_m) \qquad \text{with} \quad L(\theta) = -\frac{1}{2}\,\mathbb{E}[q_0(\theta)].$$

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

$$L(\theta^*) - L(\theta) = \frac{1}{2}\,\mathbb{E}\Big[\mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big]\Big]. \tag{3.4.5}$$

But,

$$\begin{aligned}
\mathbb{E}\big[q_0(\theta) - q_0(\theta^*) \mid \mathcal{F}_0\big] &= \mathbb{E}\Big[\frac{(X_0 - f^0_\theta)^2}{H^0_\theta} + \log(H^0_\theta) - \frac{(X_0 - f^0_{\theta^*})^2}{H^0_{\theta^*}} - \log(H^0_{\theta^*}) \mid \mathcal{F}_0\Big] \\
&= \log\Big(\frac{H^0_\theta}{H^0_{\theta^*}}\Big) + \frac{\mathbb{E}\big[(X_0 - f^0_\theta)^2 \mid \mathcal{F}_0\big]}{H^0_\theta} - \frac{\mathbb{E}\big[(X_0 - f^0_{\theta^*})^2 \mid \mathcal{F}_0\big]}{H^0_{\theta^*}} \\
&= \log\Big(\frac{H^0_\theta}{H^0_{\theta^*}}\Big) - 1 + \frac{\mathbb{E}\big[(X_0 - f^0_{\theta^*} + f^0_{\theta^*} - f^0_\theta)^2 \mid \mathcal{F}_0\big]}{H^0_\theta} \\
&= \frac{H^0_{\theta^*}}{H^0_\theta} - \log\Big(\frac{H^0_{\theta^*}}{H^0_\theta}\Big) - 1 + \frac{(f^0_{\theta^*} - f^0_\theta)^2}{H^0_\theta}
\end{aligned}$$

Thus from (3.4.5),

$$\begin{aligned}
2\,A(m) &= \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}} - \log\Big(\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}}\Big) - 1 + \frac{(f^0_{\theta^*} - f^0_{\theta^*_m})^2}{H^0_{\theta^*_m}}\Big] \\
&\geq \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}}\Big] - \log\Big(\mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}}\Big]\Big) - 1 + \mathbb{E}\Big[\frac{(f^0_{\theta^*} - f^0_{\theta^*_m})^2}{H^0_{\theta^*_m}}\Big] \quad \text{by Jensen Inequality.}
\end{aligned}$$

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

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

$$2\,A(m) = \mathbb{E}\Big[\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}} - \log\Big(\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}}\Big) - 1\Big],$$

  and from assumption **A1**, since $\theta^* \notin \Theta(m)$ and $f^0_{\theta^*} = f^0_{\theta^*_m}$, we necessarily have $H^0_{\theta^*} \neq H^0_{\theta^*_m}$ so that $\frac{H^0_{\theta^*}}{H^0_{\theta^*_m}} \neq 1$. Then $A(m) > 0$.

As a consequence,

$$\frac{C(m) - C(m^*)}{n} = A(m) + \frac{2\,c\,\kappa_n}{n}(D_m - D_{m^*}) + o_{a.s}(1).$$

That establishes (3.4.3) by virtue of (3.2.8) and as all the considered models are finite dimensional. □

In the sequel, we state and prove several lemmas to which we referred to when proving above main results.

3.4.1.c Proof of Theorem 3.3

The proof follows mutatis mutandis from the Theorem 3.1's proof after replacing line by line L_n, $\widehat{\theta}(m)$ and the criterion C by their equivalent $\widehat{L}_n$, $\widetilde{\theta}(m)$ and $\widehat{C}$ respectively and applying Lemma 3.2 instead of Proposition 3.1. □

3.4.1.d Proof of Proposition 3.1

Proof. Applying a second order Taylor expansion of L_n around $\widehat{\theta}(m)$ for n sufficiently large such that $\overline{\theta}(m) \in \Theta(m)$ which are between $\theta^*_m := (\theta^*, 0, \ldots, 0)^\top$ and $\widehat{\theta}(m)$ yields (as $\partial L_n(\widehat{\theta}(m)) = 0$ since $\widehat{\theta}(m)$ is a local extremum):

$$\begin{aligned} L_n(\theta^*) - L_n(\widehat{\theta}(m)) &= L_n(\theta^*_m) - L_n(\widehat{\theta}(m)) \\ &= \frac{1}{2}\big(\widehat{\theta}(m) - \theta^*_m\big)^\top \frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta^2}\big(\widehat{\theta}(m) - \theta^*_m\big) := I_1(m). \end{aligned}$$

From the mean value Theorem, and for large n, there exists $\overline{\theta}_{m,i}$ between $(\theta^*_m)_i$ and $(\widehat{\theta}(m))_i$ such that, $1 \leq i \leq D_m$:

$$0 = \frac{\partial L_n(\widehat{\theta}_m)}{\partial\theta_i} = \frac{\partial L_n(\theta^*_m)}{\partial\theta_i} + \frac{\partial^2 L_n(\overline{\theta}_{m,i})}{\partial\theta\partial\theta_i}(\widehat{\theta}_m - \theta^*_m) \tag{3.4.6}$$

Also, using Lemma 4 of Bardet and Wintenberger (2009) and continuous mapping Theorem, we deduce that:

$$F_n := -\Big(\frac{2}{n}\frac{\partial^2 L_n(\overline{\theta}_{m,i})}{\partial\theta\partial\theta_i}\Big)_{1\leq i\leq D_m} \xrightarrow[n\to+\infty]{a.s.} F(\theta^*_m) = \mathbb{E}\Big[\frac{\partial^2 q_0(\theta^*_m)}{\partial\theta^2}\Big]. \tag{3.4.7}$$

On the other hand, under **A2** condition, $F(\theta^*_m)$ is an invertible matrix and there exists n sufficiently large such that F_n is invertible. Therefore, from (3.4.6), it follows

$$
\begin{aligned}
\frac{I_1(m)}{2\,\log\log n} &= \frac{1}{4\,\log\log n}\big(\widehat{\theta}(m)-\theta_m^*\big)^\top\,\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta^2}\,\big(\widehat{\theta}(m)-\theta_m^*\big)\\
&= \frac{1}{4\,\log\log n}\Big(\frac{\partial L_n(\theta_m^*)}{\partial\theta}\Big)^\top\Big(-\frac{2}{n}\,F_n^{-1}\Big)\,\frac{\partial^2 L_n(\overline{\theta}(m))}{\partial\theta^2}\,\Big(-\frac{2}{n}\,F_n^{-1}\Big)\,\frac{\partial L_n(\theta_m^*)}{\partial\theta}\\
&= -\frac{1}{2n\,\log\log n}\Big(\frac{\partial L_n(\theta_m^*)}{\partial\theta}\Big)^\top\,\times\,F(\theta_m^*)^{-1}\,\times\,\frac{\partial L_n(\theta_m^*)}{\partial\theta}\,\big(1+o(1)\big)\quad\text{a.s.}
\end{aligned}
$$

The next step of the proof consists in handling the quadratic form $\Big(\frac{\partial L_n(\theta_m^*)}{\partial\theta}\Big)^\top\times F(\theta_m^*)^{-1}\times\Big(\frac{\partial L_n(\theta_m^*)}{\partial\theta}\Big)$ by applying the law of iterated logarithm (LIL).
We claim that

$$
\limsup_{n\to\infty}\frac{1}{\sqrt{2n\,\log\log n}}\,2\,G(\theta_m^*)^{-1/2}\frac{\partial L_n(\theta_m^*)}{\partial\theta}=\big(1,\ldots,1\big)^\top. \tag{3.4.8}
$$

Proof of the claim: First, since the covariance matrix of $2\,\frac{\partial L_n(\theta_m^*)}{\partial\theta}$ is $G(\theta_m^*)$, it follows that the covariance matrix of the vector $Z_m:=2\,G(\theta_m^*)^{-1/2}\,\frac{\partial L_n(\theta_m^*)}{\partial\theta}$ is the $D_m\times D_m$ identity matrix. Moreover, as

$$
\frac{\partial L_n(\theta_m^*)}{\partial\theta}=-\frac{1}{2}\sum_{t=1}^n\frac{\partial q_t(\theta_m^*)}{\partial\theta}=-\frac{1}{2}\sum_{t=1}^n\frac{\partial q_t(\theta^*)}{\partial\theta},
$$

where

$$
\mathbb{E}\Big[\frac{\partial q_t(\theta^*)}{\partial\theta}\Big|\sigma(X_{t-1},X_{t-2},\ldots,X_1)\Big]=0 \tag{3.4.9}
$$

and

$$
\mathbb{E}\Big[\Big(\frac{\partial q_1(\theta^*)}{\partial\theta_i}\Big)^2\Big]<\infty \tag{3.4.10}
$$

hold from Bardet and Wintenberger (2009) under **A4**. Finally, one can see that the i^{th} element of Z_m can be rewritten as

$$
\sum_{j=1}^{D_m}\big(2\,G(\theta_m^*)^{-1/2}\big)_{i,j}\,\frac{\partial L_n(\theta_m^*)}{\partial\theta_j}=\sum_{t=1}^n\zeta_t^i
$$

where $\zeta_t^i=-\sum_{j=1}^{D_m}\big(G(\theta_m^*)^{-1/2}\big)_{i,j}\,\frac{\partial q_t(\theta^*)}{\partial\theta_j}$. By virtue of (3.4.9), we have

$$
\mathbb{E}\Big[\zeta_t^i\Big|\sigma(X_{t-1},X_{t-2},\ldots,X_1)\Big]=0.
$$

Hence, any component of Z_m verifies the LIL. That is for any $i=1,\ldots,D_m$,

$$
\limsup_{n\to\infty}\frac{1}{\sqrt{2n\,\log\log n}}\left(2\,G(\theta_m^*)^{-1/2}\frac{\partial L_n(\theta_m^*)}{\partial\theta}\right)_i=1.
$$

This fact concludes the proof of the claim (3.4.8).

Hence writing

$$
\frac{1}{2n\,\log\log n}\frac{\partial L_n(\theta_m^*)}{\partial\theta}^\top F(\theta_m^*)^{-1}\frac{\partial L_n(\theta_m^*)}{\partial\theta}=
$$

$$
\left(\frac{1}{\sqrt{2n\,\log\log n}}\,2G(\theta_m^*)^{-1/2}\frac{\partial L_n(\theta_m^*)}{\partial\theta}\right)^\top\frac{G(\theta_m^*)^{1/2}F(\theta_m^*)^{-1}G(\theta_m^*)^{1/2}}{4}\left(\frac{1}{\sqrt{2n\,\log\log n}}\,2G(\theta_m^*)^{-1/2}\frac{\partial L_n(\theta_m^*)}{\partial\theta}\right)
$$

it follows

$$\limsup_{n\to\infty} \frac{L_n\big(\widehat{\theta}(m)\big) - L_n(\theta^*_m)}{2\,\log\log n} = \mathbf{1}_m^\top\, \Sigma_{\theta^*_m}\, \mathbf{1}_m.$$

□

3.4.1.e Proof of Proposition 3.2

It is sufficient to show that

$$\Sigma_{\theta^*_m} := G(\theta^*_m)^{1/2} F(\theta^*_m)^{-1} G(\theta^*_m)^{1/2} = \begin{pmatrix} 2\,\mathrm{I}_{D^1_m, D^1_m} & \mathrm{O}_{D^1_m, D^2_m} \\ \mathrm{O}_{D^2_m, D^1_m} & (\mu_4 - 1)\,\mathrm{I}_{D^2_m, D^2_m} \end{pmatrix}, \tag{3.4.11}$$

where I is the identity matrix and O the null matrix. From Bardet and Wintenberger (2009), we have for $m^* \subset m$ and $i, j \in m$:

$$\begin{aligned} \big(G(\theta^*_m)\big)_{i,j} &= \mathbb{E}\Big[\frac{\partial q_0(\theta^*_m)}{\partial \theta_i}\frac{\partial q_0(\theta^*_m)}{\partial \theta_j}\Big] = \mathbb{E}\Big[4(H^0_{\theta^*_m})^{-1}\,\frac{\partial f^0_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial f^0_{\theta^*_m}}{\partial \theta_j} + (\mu_4 - 1)\,(H^0_{\theta^*_m})^{-2}\,\frac{\partial H^0_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial H^0_{\theta^*_m}}{\partial \theta_j}\Big] \\ \big(F(\theta^*_m)\big)_{i,j} &= \mathbb{E}\Big[\frac{\partial^2 q_0(\theta^*_m)}{\partial \theta_i \partial \theta_j}\Big] = \mathbb{E}\Big[2(H^0_{\theta^*_m})^{-1}\,\frac{\partial f^0_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial f^0_{\theta^*_m}}{\partial \theta_j} + (H^0_{\theta^*_m})^{-2}\,\frac{\partial H^0_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial H^0_{\theta^*_m}}{\partial \theta_j}\Big] \end{aligned} \tag{3.4.12}$$

1/ If $\mu_4 = 3$, then $G(\theta^*_m) = 2\,F(\theta^*_m)$ and the result is straightforward.

2/ For the second configuration: from (4.6.1), it is clear that all the terms $F(\theta^*_m)_{i,j}$ and $G(\theta^*_m)_{i,j}$ are equals to zero for $i = 1, \ldots, D^1_m$ and $j = 1, \cdots, D^2_m$ implying

$$F(\theta^*_m) = \begin{pmatrix} A_{1,D^1_m} & \mathrm{O}_{D^1_m, D^2_m} \\ \mathrm{O}_{D^2_m, D^1_m} & B_{D^1_m+1, D^1_m+D^2_m} \end{pmatrix} \text{ and } \quad G(\theta^*_m) = \begin{pmatrix} 2\,A_{1,D^1_m} & \mathrm{O}_{D^1_m, D^2_m} \\ \mathrm{O}_{D^2_m, D^1_m} & (\mu_4 - 1)\,B_{D^1_m+1, D^1_m+D^2_m} \end{pmatrix}$$

and

$$A_{1,D^1_m} = \left(\mathbb{E}\Big[2(H^t_{\theta^*_m})^{-1}\,\frac{\partial f^t_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial f^t_{\theta^*_m}}{\partial \theta_j}\Big]\right)_{1 \le i,j \le D^1_m}$$

$$B_{D^1_m+1, D^1_m+D^2_m} = \left(\mathbb{E}\Big[(H^t_{\theta^*_m})^{-2}\,\frac{\partial H^t_{\theta^*_m}}{\partial \theta_i}\,\frac{\partial H^t_{\theta^*_m}}{\partial \theta_j}\Big]\right)_{D^1_m+1 \le i,j \le D^1_m+D^2_m}.$$

As a consequence,

$$\begin{aligned} G(\theta^*_m)\,F(\theta^*_m)^{-1} &= \mathrm{Diag}(2\,A_{1,D^1_m}, (\mu_4 - 1)\,B_{D^1_m+1, D^1_m+D^2_m})\,\mathrm{Diag}(A^{-1}_{1,D^1_m}, B^{-1}_{D^1_m+1, D^1_m+D^2_m}) \\ &= \mathrm{Diag}(2\,\mathrm{I}_{p_1}, (\mu_4 - 1)\,\mathrm{I}_{D^2_m}). \end{aligned}$$

As a covariance matrix, $G(\theta^*_m)$ is positive definite. Therefore the square root $G(\theta^*_m)^{1/2}$ of $G(\theta^*_m)$ is unique and blocks diagonal. Thus,

$$\begin{aligned} \Sigma_{\theta^*_m} &= G(\theta^*_m)^{-1/2}\Big(G(\theta^*_m)F(\theta^*_m)^{-1}\Big)G(\theta^*_m)^{1/2} \\ &= G(\theta^*_m)F(\theta^*_m)^{-1}, \end{aligned}$$

which gives (3.4.11). □

3.4.2 Technical Lemmas

Lemma 3.1. *Let $X \in \mathcal{H}(M_{\theta^*}, f_{\theta^*})$ (or $\widetilde{\mathcal{H}}(H_{\theta^*})$) and $\Theta \subseteq \Theta(r)$ (or $\Theta \subseteq \widetilde{\Theta}(r)$) with $r \geq 4$. Assume that assumption **A3** holds. Then for $i = 0, 1, 2$, it holds*

$$\frac{1}{\log\log n} \left\| \frac{\partial^{(i)} \widehat{L}_n(\theta)}{\partial \theta^i} - \frac{\partial^{(i)} L_n(\theta)}{\partial \theta^i} \right\|_\Theta \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} 0. \tag{3.4.13}$$

Proof. This Lemma has already been proved in Bardet et al. (2020b) in a more general framework. Let us prove the result for $i = 0$. Other cases can be deduced by using a similar reasoning.

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

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

By Corollary 1 of Kounias and Weng (1969), (3.4.13) is established when:

$$\sum_{k\geq 1} \frac{1}{\log\log k} \mathbb{E}\left\|\widehat{q}_k(\theta) - q_k(\theta)\right\|_\Theta < \infty. \tag{3.4.14}$$

From Bardet and Wintenberger (2009) and Bardet et al. (2020b), there exists a constant C such that
1/ If $X \in \mathcal{H}(M_\theta, f_\theta)$, we deduce

$$\mathbb{E}\big[\|(\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta\big] \leq C \Big(\sum_{j\geq t} \alpha_j(f_\theta, \Theta) + \sum_{j\geq t} \alpha_j(M_\theta, \Theta) \Big). \tag{3.4.15}$$

Hence,

$$\sum_{k\geq 1} \frac{1}{\log\log k} \mathbb{E}\big[\|\widehat{q}_k(\theta) - q_k(\theta)\|_\Theta\big] \leq C \sum_{k\geq 1} \frac{1}{\log\log k} \Big(\sum_{j\geq k} \alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) \Big),$$

which is finite by definition of the class $\mathcal{H}(M_\theta, f_\theta)$, and this achieves the proof.

2/ If $X \in \widetilde{\mathcal{H}}(\widetilde{H}_\theta)$,

$$\mathbb{E}\big[\|(\widehat{q}_t(\theta) - q_t(\theta)\|_\Theta\big] \leq C \Big(\sum_{j\geq t} \alpha_j(H_\theta, \Theta) \Big). \tag{3.4.16}$$

This fact along with Corollary 1 of Kounias and Weng (1969) enable us to conclude the proof in this case. □

Lemma 3.2. *Under the assumptions of Theorem 3.3, for any model $m \in \mathcal{M}$ with $\theta^* \in \overset{\circ}{\widetilde{\Theta(m)}}$, it holds*

$$\limsup_{n\to\infty} \left(\frac{\widehat{L}_n\big(\widetilde{\theta}(m)\big) - \widehat{L}_n(\theta^*)}{2\,\log\log n} \right) \quad = \quad c_{min}(m) \quad a.s. \tag{3.4.17}$$

Proof. Applying a second order Taylor expansion of $\widetilde{L}_n$ around $\widetilde{\theta}(m)$ for n sufficiently large such that $\overline{\theta}(m) \in \Theta(m)$ which are between $\theta^*_m := (\theta^*, 0, \ldots, 0)^\top$ and $\widetilde{\theta}(m)$ yields (as $\partial \widehat{L}_n(\widetilde{\theta}(m)) = 0$ since $\widetilde{\theta}(m)$ is a local extremum):

$$\begin{aligned}
\widehat{L}_n(\theta^*) - \widehat{L}_n\big(\widetilde{\theta}(m)\big) &= \widehat{L}_n(\theta^*_m) - \widehat{L}_n\big(\widetilde{\theta}(m)\big) \\
&= \frac{1}{2}\big(\widetilde{\theta}(m) - \theta^*_m\big)^\top \frac{\partial^2 \widehat{L}_n(\overline{\theta}(m))}{\partial \theta^2} \big(\widetilde{\theta}(m) - \theta^*_m\big) \\
&:= I_2(m).
\end{aligned}$$

But $I_2(m)$ can be rewritten as

$$\begin{aligned}
\frac{I_2(m)}{2\,\log\log n} = \frac{1}{4}\big(\widetilde{\theta}(m) - \theta^*_m\big)^\top \frac{1}{\log\log n}\left[\frac{\partial^2 \widehat{L}_n(\overline{\theta}(m))}{\partial \theta^2} - \frac{\partial^2 L_n(\overline{\theta}(m))}{\partial \theta^2}\right]\big(\widetilde{\theta}(m) - \theta^*_m\big) \\
+ \frac{1}{4}\frac{1}{\log\log n}\big(\widetilde{\theta}(m) - \theta^*_m\big)^\top \frac{\partial^2 L_n(\overline{\theta}(m))}{\partial \theta^2}\big(\widetilde{\theta}(m) - \theta^*_m\big) \\
=: I_{21}(m) + I_{22}(m).
\end{aligned}$$

First, as $\widetilde{\theta}(m) \xrightarrow[n\to+\infty]{a.s.} \theta^*_m$ along side with Lemma 3.1, it follows

$$I_{21}(m) \xrightarrow[n\to+\infty]{a.s.} 0. \tag{3.4.18}$$

Writing a counterpart of (3.4.6) using the quasi-functions, we have

$$\widetilde{\theta}(m) - \theta^*_m = \Big(\partial^2 \widehat{L}_n(\overline{\theta}(m))\Big)^{-1} \frac{\partial \widehat{L}_n(\theta^*_m)}{\partial \theta}$$

Hence $I_{22}(m)$ can be rewritten as

$$\begin{aligned}
I_{22}(m) &= \frac{1}{4\,\log\log n}\big(\widetilde{\theta}(m) - \theta^*_m\big)^\top \frac{\partial^2 L_n(\overline{\theta}(m))}{\partial \theta^2}\big(\widetilde{\theta}(m) - \theta^*_m\big) \\
&= \frac{1}{4\,\log\log n}\Big(\frac{\partial \widehat{L}_n(\theta^*_m)}{\partial \theta}\Big)^\top \Big(\partial^2 \widehat{L}_n(\overline{\theta}(m))\Big)^{-1} \Big(\partial^2 L_n(\overline{\theta}(m))\Big) \Big(\partial^2 \widehat{L}_n(\overline{\theta}(m))\Big)^{-1} \frac{\partial \widehat{L}_n(\theta^*_m)}{\partial \theta} \\
&= -\frac{1}{2\,n\,\log\log n}\Big(\frac{\partial L_n(\theta^*_m)}{\partial \theta}\Big)^\top \times F(\theta^*_m)^{-1} \times \frac{\partial L_n(\theta^*_m)}{\partial \theta}\big(1 + o(1)\big) \quad \text{a.s.}
\end{aligned} \tag{3.4.19}$$

since from (3.4.7), it holds

$$-\frac{n}{2}\Big(\partial^2 L_n(\overline{\theta}(m))\Big)^{-1} \xrightarrow[n\to+\infty]{a.s.} F(\theta^*_m)^{-1}$$

and along with Lemma 3.1, it also holds

$$-\frac{n}{2}\log\log n\Big(\partial^2 \widehat{L}_n(\overline{\theta}(m))\Big)^{-1} \xrightarrow[n\to+\infty]{a.s.} F(\theta^*_m)^{-1}.$$

As a consequence, the chain of following equalities holds a.s.

$$\begin{aligned}
\limsup_{n\to\infty}\left(\frac{\widehat{L}_n\big(\widetilde{\theta}(m)\big) - \widehat{L}_n(\theta^*)}{2\,\log\log n}\right) &= \limsup_{n\to\infty}\left(-\frac{1}{2\,n\,\log\log n}\Big(\frac{\partial L_n(\theta^*_m)}{\partial \theta}\Big)^\top \times F(\theta^*_m)^{-1} \times \frac{\partial L_n(\theta^*_m)}{\partial \theta}\right) \\
&= c_{min}(m)
\end{aligned}$$

That ends the proof of (3.4.17). □

4

Efficient and consistent data-driven model selection for time series

The content of this chapter is based on a work in progress in collaboration with J-M BARDET and W. KENGNE.

Abstract

This chapter 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. On the one hand, by seeking a penalty that minimizes the risk induced by the quasi-likelihood, we establish the asymptotic efficiency of a data-driven penalty generalizing the AIC penalty term. On the other hand, by a Bayesian approach, we also show the asymptotic consistency of a data-driven penalty that generalizes the BIC criterion. Monte Carlo experiments are performed to highlight the obtained results.

4.1 Introduction

Model selection is one of the fundamental tasks in scientific research specially in Statistics and Data Science. It aims at providing a model (or an algorithm) that is best, following a criterion, suitable to observed data in order to gain insight on the underlying data generating process or/and to make good forecasts.

Two leading model selection procedures have received a lot of attention in the literature. On one hand, the resampling methods such as hold out or more generally V-fold cross-validation are widely used in machine learning community. On the other hand, the methods of penalization are also now very popular in the community of applied or theoretical statisticians..

The main challenging task when designing a penalized based criterion is the calibration of the penalty. This is mainly dependent on the goal one would like the final criterion achieves. For instance, the objective could be the *consistency*, the *efficiency* or the *adaptive nature in the minimax sense* to name few.

The consistency property aims at identifying the data generating process with high probability. Hence, its requires the assumption whereby there exists a true model in the set of competitive models and the goal is to select this with probability approaches one as the sample size tends to infinity.

Although the consistency is a convincing mathematical property, this asymptotic property is not always the most interesting when switching to a practical implementation. Indeed the true underlying process is generally unknown and trying to identify the true model for any data is quite ambitious. It is often more plausible to assume that the true data generating process is infinite-dimensional, and that one tries to identify a "good" finite-dimensional model based on the data (Hurvich and Tsai (1989)). Therefore, it is common in this framework to let the dimension of the competitive models to depend on the number of observations in order to obtain better approximation and to reduce the risk of prediction. Hence, the model selection is said to be efficient when its risk is asymptotically equivalent to the risk of the *oracle*.

In this work, we are interested by providing efficient and consistent penalized data-driven criteria for affine causal times series.

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

$$X_t = M\big((X_{t-i})_{i\in\mathbb{N}^*}\big)\,\xi_t + f\big((X_{t-i})_{i\in\mathbb{N}^*}\big) \text{ for any } t \in \mathbb{Z}. \tag{4.1.1}$$

where $(\xi_t)_{t\in T}$ is a sequence of zero-mean independent identically distributed random vectors (i.i.d.r.v) satisfying $\mathbb{E}(|\xi_0|^r) < \infty$ with $r \geq 1$ and M, $f : \mathbb{R}^\infty \to \mathbb{R}$ are two measurable functions, where $\mathbb{R}^\infty$ is the set of numeric sequence with finite number of non-zero terms.

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.

Note that, numerous classical time series models such as ARMA(p,q), GARCH(p,q), ARMA(p,q)-GARCH(p,q) (see Ding et al. (1993) and Ling and McAleer (2003)) or APARCH(δ,p,q) processes (see Ding et al. (1993)) belongs to $\mathcal{AC}(M,f)$.
The study of this type of process more often requires the classical regularity conditions on the functions M and f that are not really restrictive and remain valid for many time series.

In this semi-parametric framework, we consider $(f_\theta)_{\theta\in\Theta}$ and $(M_\theta)_{\theta\in\Theta}$ two families of functions such as for $\theta \in \Theta$, $f_\theta : \mathbb{R}^\infty \to \mathbb{R}$ and $M_\theta : \mathbb{R}^\infty \to [0,\infty)$ are known.

There already exist several important contributions devoted to the model selection for time series ; we refer to the book of McQuarrie and Tsai (1998) and the references therein for an overview on this topic. Also, the time series model selection literature is very extensive and still growing ; we refer to the monograph of Rao et al. (2001), 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 Ding et al. (2018).

The asymptotically efficient selection property have already been tackled: by Shibata (1980), and recently by Hsu et al. (2019b)). But this was done for linear process type $AR(\infty)$. In this paper, we focus on the class of models (4.1.1), and addressed this questions: what regularity conditions are sufficient to build a efficient model selection criteria? Are these obtained criteria data-driven? Is the classic criterion such as AIC, still have efficient property for choosing a model among the collection $\mathcal{M}$?
These questions have not yet been answered for the affine causal class and the framework considered here. This new contribution provides theoretical and numerical response of these issues. Also, following the derivation technique of the BIC criterion, we propose a data driven criterion allowing to obtain better results in consistency.
The paper is organized as follows. The model selection framework along with notations and assumptions are described in Section 4.2. Efficient criteria and the asymptotic efficiency are studied in Section 4.4. In Section 4.5, the derivation of a consistent data-driven BIC type criteria for affine causal class is studied. Section 4.7 provides some simulation experiments and finally, Section 4.8 contains the proofs.

4.2 Model selection framework

4.2.1 Finite family $\mathcal{M}$ of parametric affine causal models

Assume a trajectory $(X_1,\ldots,X_n)$ is observed from a causal stationary solution of (4.1.1) where M and f are two known functions indexed by an unknown finite dimensional vector of parameters θ^*. Also $(X_1,\ldots,X_n)$ is supposed to be sampled according to a joint

distribution $\mathbb{P}_{(X_1,\dots,X_n)}$.

Now consider a finite family $\mathcal{M}$ of models belonging to parametric affine causal models. In Proposition 1 of Bardet et al. (2020b), due to the linearity of such models, it was established that it is always possible to find a dimension $d \in \mathbb{N}^*$ and a unique couple of known functions (M_θ, f_θ) with $\theta \in \mathbb{R}^d$ in such a way that any model $m \in \mathcal{M}$ belongs to the class $\mathcal{AC}(M_\theta, f_\theta)$. More precisely, there is a one-to-one correspondence between each model $m \in \mathcal{M}$ and a linear subspace $\Theta_m \subset \mathbb{R}^d$ and $\dim(\Theta_m) = |m|$ the number of unknown parameters of the model m. As a consequence, if we denote m^* the "true" model corresponding to $\mathcal{AC}(M_{\theta^*}, f_{\theta^*})$, we will say:

- if $m \in \mathcal{M}$ is such that $\Theta_{m^*} \subset \Theta_m$ and $\Theta_{m^*} \neq \Theta_m$ (also denoted $m^* \subset m$ and $m^* \neq m$), this is an overfitting's case;
- if $m \in \mathcal{M}$ is such that $\Theta_{m^*} \not\subset \Theta_m$ (also denoted $m^* \not\subset m$) , this is a misspecified case.

For example, if m^* corresponds to a AR(2) process and if $\mathcal{M}$ contains AR($p_{\max}$) processes and ARCH($q_{\max}$), we have $d = 1 + p_{\max} + q_{\max}$ and for $\theta = (\theta_i)_{0\le i\le d}$,

$$f_\theta((X_{t-k})_{k\ge1}) = \sum_{i=1}^{p_{\max}} \theta_i\, X_{t-i} \qquad \text{and} \qquad M_\theta((X_{t-k})_{k\ge1}) = \Big(\theta_0 + \sum_{i=p_{\max}+1}^{p_{\max}+q_{\max}} \theta_i\, X_{t-i}^2\Big)^{1/2}.$$

Then $\Theta_{m^*} = \big\{(\theta_0,\theta_1,\theta_2,0,\dots,0),\ (\theta_0,\theta_1,\theta_2) \in \mathbb{R}^3\big\}$, an AR(4) process implies an overfitting, while an AR(1) or an ARCH(2) process implies a misspecified case.
In the sequel, we will always assume that

$$m^* \in \mathcal{M}.$$

But m^* is supposed to be unknown and our goal is to find a "best" model among a finite family $\mathcal{M}$ that forecasts with a minimum risk (defined in Subsection 4.2.2) or the most probable model after observing the trajectory $(X_1,\dots,X_n)$ (see Section 4.5). This section aims at describing the Maximum and Quasi-Maximum Likelihood contrasts that provides our chosen risk and presenting a data-driven penalization procedure for model selection.

4.2.2 Maximum Likelihood Estimation

Given a predictor θ, we measure its quality by the risk defined as

$$R(\theta) := \mathbb{P}\gamma(\theta) = \mathbb{E}[\gamma(\theta, X_1)]$$
$$\text{with} \quad \gamma(\theta, X_t) := \frac{(X_t - f_\theta^t)^2}{H_\theta^t} + \log(H_\theta^t), \text{and} \quad \left\{ \begin{array}{rcl} f_\theta^t & := & f_\theta\big((X_{t-k})_{k\ge1}\big) \\ M_\theta^t & := & M_\theta\big((X_{t-k})_{k\ge1}\big) \\ H_\theta^t & := & \big(M_\theta^t\big)^2 \end{array} \right. \tag{4.2.1}$$

By referring to Massart (2007) or Francq and Zakoian (2010), the contrast $\gamma(\theta,.)$ is -2 times the Gaussian conditional log-density of X_t. Moreover, the Gaussian MLE is derived from the conditional (with respect to the filtration $\sigma\big\{(X_t)_{t\le0}\big\}$) log-likelihood of $(X_1,\dots,X_n)$ when (ξ_t) is supposed to be a Gaussian standard white noise. We deduce that this conditional log-likelihood (up to an additional constant) L_n is defined for a parameter θ by:

$$L_n(\theta) := -\frac{1}{2}\sum_{t=1}^{n} \gamma(\theta, X_t) \tag{4.2.2}$$

As proved in Bardet and Wintenberger (2009), the risk function R achieves its unique minimum at the "true"' parameter θ^* over any parameter set Θ, when $\theta^* \in \Theta$

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, R(\theta). \tag{4.2.3}$$

Therefore θ^* is considered as an ideal for the model selection procedure and serves as a benchmark to compare predictors. Given a model $m \in \mathcal{M}$ with Θ_m its parameter space that does not necessary contains θ^*, let define

$$\theta_m^* = \underset{\theta \in \Theta_m}{\operatorname{argmin}}\, R(\theta). \tag{4.2.4}$$

As a consequence, we have:

$$\theta_{m^*}^* = \theta^* \quad \text{and more generally, if } m^* \subset m, \quad \theta_m^* = \theta^*.$$

Besides of minimizing the risk, we also consider the minimization of the natural associated loss function, which is defined as

$$\ell(\theta, \theta^*) := R(\theta) - R(\theta^*) \geq 0. \tag{4.2.5}$$

This is a well-known measure of separation between the candidate model generated by θ and the true one indexed by θ^*.

Let us introduce the empirical distribution $\mathbb{P}_n$ be such that

$$\mathbb{P}_n = \frac{1}{n} \sum_{t=1}^{n} \delta_{X_t},$$

where δ_{X_t} is the Dirac distribution at observation X_t. Let set by γ_n the associated empirical criterion

$$\gamma_n(\theta) := \mathbb{P}_n \gamma(\theta, .) = \frac{1}{n} \sum_{t=1}^{n} \gamma(\theta, X_t).$$

We have $\gamma_n(\theta) = -\frac{2}{n} L_n(\theta)$, so that maximizing the log-likelihood is equivalent to minimize the empirical criterion γ_n.

4.2.3 Quasi-Maximum Likelihood Estimation

Since the the white noise is not necessary a Gaussian one and since the log-likelihood (and then the empirical risk) $L_n(\theta)$ depends on $(X_t)_{t \leq 0}$ that are unknown, a quasi-log-likelihood can be used as an approximation of the log-likelihood. It It consists of replacing $\gamma(\theta, X_t)$ by an approximation $\widehat{\gamma}(\theta, X_t)$. Hence, in our framework, we consider the following conditional quasi log-likelihood (up to an additional constant) given for all $\theta \in \Theta$ by

$$\widehat{L}_n(\theta) := -\frac{1}{2} \sum_{t=1}^{n} \widehat{\gamma}(\theta, X_t)$$

$$\text{with } \widehat{\gamma}(\theta, X_t) := \frac{(X_t - \widehat{f}_\theta^t)^2}{\widehat{H}_\theta^t} + \log(\widehat{H}_\theta^t) \quad \text{and} \quad \begin{cases} \widehat{f}_\theta^t & := & f_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{M}_\theta^t & := & M_\theta(X_{t-1}, X_{t-2}, \cdots, X_1, u) \\ \widehat{H}_\theta^t & := & (\widehat{M}_\theta^t)^2 \end{cases} \tag{4.2.6}$$

for any deterministic sequence $u = (u_n)_{n\in\mathbb{N}}$ with finitely many non-zero values (we will use $u = 0$ without loss of generality).
In addition the computable empirical risk is then:

$$\widehat{\gamma}_n(\theta) = \mathbb{P}_n\widehat{\gamma}(\theta, .) = -\frac{2}{n}\widehat{L}_n(\theta).$$

Finally, for each specific model $m \in \mathcal{M}_n$, we define the Gaussian Quasi-Maximum Likelihood Estimator (QMLE) $\widehat{\theta}_m$ as

$$\widehat{\theta}_m = \underset{\theta\in\Theta_m}{\operatorname{argmin}}\ \widehat{\gamma}_n(\theta). \tag{4.2.7}$$

The estimator $\widehat{\theta}_m$ is commonly called the *Empirical Risk Minimizer* (ERM).

4.2.4 The penalization procedure

For $m \in \mathcal{M}$, the ERM provides an estimator in Θ_m. The goal is to come up with the model that minimizes the excess loss over $\mathcal{M}$

$$\inf_{m\in\mathcal{M}} \ell(\widehat{\theta}_m, \theta^*). \tag{4.2.8}$$

This model is unknown since (4.2.8) depends on θ^* and the distribution $P_{(X_1,\dots,X_n)}$ that are unknown.

A classical way to solve (4.2.8) problem is to design for every $m \in \mathcal{M}$ an estimator of $R(\widehat{\theta}_m)$ and we naturally choose s $\widehat{\gamma}_n(\widehat{\theta}_m)$. First, it is well known that the empirical criterion $\widehat{\gamma}_n(\widehat{\theta}_m)$ is an optimist version of $R(\widehat{\theta}_m)$ and decreases with the dimension of the model. Therefore, it is common to add a penalty term to counteract this bias. As a consequence, define a function pen, which is called the penalty function, possibly data-dependent, such as pen: $m \in \mathcal{M} \mapsto \mathrm{pen}(m) \in \mathbb{R}^+$. Then define the penalized contrast and the model selected by it:

$$\widehat{m}_{\mathrm{pen}} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{\widehat{C}_{\mathrm{pen}}(m)\right\} \qquad \text{with} \qquad \widehat{C}_{\mathrm{pen}}(m) := \widehat{\gamma}_n\big(\widehat{\theta}_m\big) + \mathrm{pen}(m). \tag{4.2.9}$$

In order to achieve (4.2.8), the *ideal penalty* to consider in (4.2.9) is

$$\mathrm{pen}_{id}(m) = R(\widehat{\theta}_m) - \widehat{\gamma}_n(\widehat{\theta}_m). \tag{4.2.10}$$

Using its definition we have:

$$\widehat{m}_{id} := \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{\ell(\widehat{\theta}_m, \theta^*)\right\} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{R(\widehat{\theta}_m)\right\} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{\widehat{C}_{\mathrm{pen}_{id}}(m)\right\}. \tag{4.2.11}$$

However, the function R is unknown (as it depends on the unknown distribution of the process) therefore pen_{id} cannot generally be used directly.
The question is how to choose the penalty in (4.2.9) so that $\widehat{m}_{\mathrm{pen}}$ mimics the *oracle*, which is the model associated with the minimum risk *i.e.* (4.2.8). Hence, we would like our final estimator $\widehat{\theta}_{\widehat{m}_{\mathrm{pen}}}$ to behave asymptotically like the oracle. That is to satisfy:

$$\mathbb{P}\Big(\ell(\widehat{\theta}_{\widehat{m}_{\mathrm{pen}}}, \theta^*) \le \min_{m\in\mathcal{M}} \left\{\ell(\widehat{\theta}_m, \theta^*)\right\} + \frac{C}{n}\Big) \underset{n\to\infty}{\longrightarrow} 1 \tag{4.2.12}$$

and/or for any $n \geq n_0$

$$\mathbb{E}\big[\ell(\widehat{\theta}_{\widehat{m}_{\text{pen}}}, \theta^*)\big] \leq \min_{m \in \mathcal{M}} \Big\{ \mathbb{E}\big[\ell(\widehat{\theta}_m, \theta^*)\big] \Big\} + \frac{C}{n}. \tag{4.2.13}$$

The aim of this paper is to find a good choice of pen(m) in order to obtain the *asymptotic optimality* (4.2.13), that we prefer to (4.2.12), which means that the expectation of the risk of the chosen estimator is equivalent to the expectation of the risk of the oracle when the sample size tends to infinity.

4.3 Asymptotic behavior of the QMLE

Before considering the problem of model selection, we establish a central limit theorem satisfied by $\widehat{\theta}_m$) for any model $m \in \mathcal{M}$, *i.e.* as well if m is an overfitted or a misspecified model. Before this, some notations and assumptions have to be precised.

4.3.1 Notations and main assumptions

In the sequel, we will consider a subset Θ of $\mathbb{R}^d$ which is compact. We will use the following norms:

- $\|.\|$ denotes the usual Euclidean norm on $\mathbb{R}^\nu$, with $\nu \geq 1$;
- for a matrix A, denote $\|A\|$ the subordinate matrix norm such that $\|A\| = \sup_{v \neq 0} \frac{\|A\, v\|}{\|v\|}$;
- if X is a $\mathbb{R}^\nu$-random variable and $r \geq 1$, we set $\|X\|_r = \big(\mathbb{E}\big[\|X\|^r\big]\big)^{1/r} \in [0, \infty]$;
- for $\theta \in \Theta \subset \mathbb{R}^d$, if $\Psi_\theta : \mathbb{R}^\infty \to E$ where $E = \mathbb{R}^\nu$ or E is a set of square matrix, denote $\|\Psi_\theta(\cdot)\|_\Theta = \sup_{\theta \in \Theta} \big\{ \|\Psi_\theta(\cdot)\| \big\}$;
- for $\theta \in \Theta \subset \mathbb{R}^d$, if $\Psi_\theta : \mathbb{R}^\infty \to \mathbb{R}$ is a $\mathcal{C}^2(\Theta \times \mathbb{R}^\infty)$ function, we will denote

  $$\partial_\theta \Psi_\theta(\cdot) = \Big(\frac{\partial}{\partial \theta_i} \Psi_\theta(\cdot) \Big)_{1 \leq i \leq d} = \big(\partial_{\theta_i} \Psi_\theta(\cdot) \big)_{1 \leq i \leq d} \quad \text{and} \quad \partial^2_{\theta^2} \Psi_\theta(\cdot) = \Big(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \Psi_\theta(\cdot) \Big)_{1 \leq i,j \leq d};$$

- consider $\Psi_\theta : \mathbb{R}^\infty \to \mathbb{R}$ for any $\theta \in \Theta \subset \mathbb{R}^d$. Then, we define:

  $\mathbf{A}(\Psi_\theta, \Theta)$: $\|\Psi_\theta(0)\|_\Theta < \infty$ 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:

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

Several assumptions on the AC class will be considered thereafter:

Assumption A0: *The process $X \in \mathcal{AC}(M_{\theta^*}, f_{\theta^*})$ where $\theta^* \in \Theta$ is defined in* (4.1.1) *where:*

- *the white noise $(\xi_t)_t$ is such as $\|\xi_0\|_r < \infty$ with $8 < r$;*
- *for any $x \in \mathbb{R}^\infty$, the functions $\theta \to M_\theta$ and $\theta \to f_\theta$ are $\mathcal{C}^2(\Theta)$ functions:*

- $\Theta \in \mathbb{R}^d$ *is a compact set such as*

$$\Theta \subset \Big\{\theta \in \mathbb{R}^d,\ A(f_\theta, \{\theta\}) \text{ and } A(M_\theta, \{\theta\}) \text{ hold with } \sum_{k=1}^{\infty} \alpha_k(f_\theta, \{\theta\}) + \|\xi_0\|_r \sum_{k=1}^{\infty} \alpha_k(M_\theta, \{\theta\}) < 1\Big\}. \quad (4.3.1)$$

Under this assumption, Doukhan and Wintenberger (2008) showed that there exists a stationary causal (*i.e.* X_t is depending only on $(X_{t-k})_{k\in\mathbb{N}}$ for any $t \in \mathbb{Z}$) and ergodic solution of (4.1.1) with 8-order moment for any $\theta \in \Theta$.

Now assumption **A0** holds. We will also add several assumptions required for insuring the strong consistency and the asymptotic normality of the QMLE:

The following classical assumption ensures the identifiability of the θ^*.

Assumption A1: *For all* θ, $\theta' \in \Theta$, $(f_\theta^0 = f_{\theta'}^0$ *and* $M_\theta^0 = M_{\theta'}^0)$ *a.s.* $\implies \theta = \theta'$.

Next, the following Assumption ensures the invertibility of the "Fisher's information matrix" and is necessary to prove the asymptotic normality of the QMLE.

Assumption A2: $< \alpha, \partial_\theta f_\theta^0 >= 0 \implies \alpha = 0$ *a.s. or* $< \alpha, \partial_\theta H_\theta^0 >= 0 \implies \alpha = 0$ *a.s.*

The definition of the computable empirical risk and require that its denominators do not vanish. Hence, we are going to assume throughout this paper that the lower bound of $H_\theta(\cdot) = \big(M_\theta(\cdot)\big)^2$ is strictly positive:

Assumption A3: $\exists \underline{h} > 0$ *such that* $\inf_{\theta\in\Theta}(H_\theta(x)) \geq \underline{h}$ *for all* $x \in \mathbb{R}^\infty$.

The following assumption is a technical classical condition (see Lv and Liu (2014)).

Assumption A4: *For every* $m \in \mathcal{M}$, *if* $(\overline{\theta}_{m,n})$ *is a sequence of* Θ_m *satisfying* $\overline{\theta}_{m,n} \xrightarrow[n\to+\infty]{a.s.} \theta^*$, *then*

$$\limsup_{n\to\infty} \Big\{\mathbb{E}\Big[\Big(\Big\|\frac{1}{n}\big(\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_m)\big)_{i,j\in m}\Big)^{-1}\Big\|^8\Big]\Big\} < \infty. \quad (4.3.2)$$

Remark 4.1. *Note that under assumption* **A0**, *if* $\overline{\theta}_{m,n} \xrightarrow[n\to+\infty]{a.s.} \theta_m^*$ *then*

$$\Big\|\Big(\frac{1}{n}\big(\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_{m,n})\big)_{i,j\in m}\Big)^{-1}\Big\|^8 \xrightarrow[n\to+\infty]{a.s.} \Big\|\Big(\big(-\frac{1}{2}\,\partial^2_{\theta_i\theta_j}\gamma(\theta_m^*)\big)_{i,j\in m}\Big)^{-1}\Big\|^8$$

Thus, from the Egorov's Theorem, we can find an event $\widetilde{\Omega}$ *with sufficiently large probability such that the relation (4.3.2) in the assumption* **A4** *holds if the expectation is taken on the event* $\widetilde{\Omega}$. *For the particular case of the linear processes, the assumption* **A4** *holds true under a mild condition on the distribution of* X, *see for instance Papangelou (1994) and Findley and Wei (2002).*

Finally, the decrease rates of $(\alpha_j(f_\theta, \Theta))_j$, $(\alpha_j(M_\theta, \Theta))_j$, $(\alpha_j(\partial_\theta f_\theta, \Theta))_j$ and $(\alpha_j(\partial_\theta M_\theta, \Theta))_j$ have to be fast enough for insuring the strong consistency and the asymptotic normality of the QMLE:

Assumption A5: *Conditions* $\mathbf{A}(f_\theta, \Theta)$, $\mathbf{A}(M_\theta, \Theta)$, $\mathbf{A}(\partial_\theta f_\theta, \Theta)$, $\mathbf{A}(\partial_\theta M_\theta, \Theta)$, $\mathbf{A}(\partial^2_{\theta^2} f_\theta, \Theta)$ *and* $\mathbf{A}(\partial^2_{\theta^2} M_\theta, \Theta)$ *hold with*

$$\alpha_j(f_\theta, \Theta) + \alpha_j(M_\theta, \Theta) + \alpha_j(\partial_\theta f_\theta, \Theta) + \alpha_j(\partial_\theta M_\theta, \Theta) = O(j^{-\delta}) \quad \text{where} \quad \delta > 7/2.$$

Note that Assumption A5 does not allow to consider long-range dependent processes, but usual short memory causal time series satisfy this assumption.

4.3.2 New asymptotic results satisfied by $\widehat{\theta}_m$

The asymptotic normality of $\widehat{\theta}_m$ has been already established in Bardet and Wintenberger (2009) when $m = m^*$ and in Bardet et al. (2020b) when $m^* \subset m$ (overfitting). This property can also be extended in the case of misspecified model, *i.e.* when $m^* \not\subset m$. Before this, we define the following key matrix for $\theta_m \in \Theta_m$ with $m \in \mathcal{M}$,

$$F(\theta_m) := \Big(\mathbb{E}\Big[\partial^2_{\theta_i \theta_j} \gamma(\theta_m, X_0) \Big] \Big)_{i,j \in m}. \tag{4.3.3}$$

First, the following corollary can be established as a particular case of more general result, Proposition 4.3, which is stated in Section 4.8 devoted to the proofs.

Corollary 4.1. *For $m \in \mathcal{M}$, let $\theta_m \in \Theta_m$ and denote*

$$\begin{aligned} G(\theta_m) &:= \frac{1}{4} \Big(\sum_{t \in \mathbb{Z}} Cov\big(\partial_{\theta_i} \gamma(\theta_m, X_0) \, , \, \partial_{\theta_j} \gamma(\theta_m, X_t) \big) \Big)_{i,j \in m} \\ \Longrightarrow \quad G(\theta^*_m) &= \frac{1}{4} \Big(Cov\big(\partial_{\theta_i} \gamma(\theta^*_m, X_0) \, , \, \partial_{\theta_j} \gamma(\theta^*_m, X_0) \big) \Big)_{i,j \in m} \quad \text{if } m^* \subset m. \end{aligned} \tag{4.3.4}$$

Then, under assumptions **A0**-**A5**, *with θ^*_m defined in* (4.2.4),

$$\frac{1}{\sqrt{n}} \Big(\partial_{\theta_j} L_n(\theta^*_m) \Big)_{j \in m} \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}\big(0 \, , \, G(\theta^*_m) \big). \tag{4.3.5}$$

Using mainly this new result, we also obtain:

Theorem 4.1. *Under assumptions* **A0**-**A5**, *for any $m \in \mathcal{M}$,*

$$\sqrt{n} \left((\widehat{\theta}_m)_i - (\theta^*_m)_i \right)_{i \in m} \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}\Big(0 \, , \, \big(F(\theta^*_m) \big)^{-1} G(\theta^*_m) \big(F(\theta^*_m) \big)^{-1} \Big), \tag{4.3.6}$$

with F defined in (4.3.3) *and G in* (4.3.4).

Hence, even in the misspecified case, $\widehat{\theta}_m$ satisfies a central limit theorem. We will use this result several times, in particular to prove that the probability of selecting a misspecified model tends quickly enough towards 0. Another technical result will also be useful for the sequel:

Proposition 4.1. *Under assumptions* **A0**-**A5**, *with $8/3 < r' \leq (8+r)/6$ and $r' < 2(\delta - 1)$ where $\delta > 7/2$ is given in Assumption A5 and for any $m \in \mathcal{M}$, then we have*

$$\sup_{n \in \mathbb{N}^*} \left\| \sqrt{n} \left((\widehat{\theta}_m)_i - (\theta^*_m)_i \right)_{i \in m} \right\|_{r'} < \infty. \tag{4.3.7}$$

Note that we also have $\sup_{n \in \mathbb{N}^*} \left\| \sqrt{n} \left((\widehat{\theta}_m)_i - (\theta^*_m)_i \right)_{i \in m} \right\|_2 < \infty$. This result will be essential for establishing the asymptotic behavior of the expectation of the ideal penalty.

4.4 Efficient model selection result

The expectation of the ideal penalty (4.2.10) has been computed (or asymptotically approximated) in several frameworks (see Mallows (1973), Akaike (1973), Schwarz (1978), Hurvich and Tsai (1989), Cavanaugh (1997); etc) and it is most often proportional to the dimension of the model (denoted D_m in the sequel).

- Mallows (1973) in the regression setting and the penalty is $2D_m\,\sigma^2/n$;
- Akaike (1973) in the density estimation framework where the penalty is D_m/n;
- Lv and Liu (2014) in misspecified density estimation and the penalty is $\mathrm{Trace}\big(B_n\,A_n^{-1}\big)/n$ where A_n is the opposite of the Hessian matrix of the log-likelihood and B_n the Fisher Information matrix.

In order to approximate (4.2.10) in this framework, let first provide a decomposition of this term in order to facilitate the computation. For any model $m \in \mathcal{M}$, write

$$\mathrm{pen}_{id}(m) := R(\widehat{\theta}_m) - \widehat{\gamma}_n(\widehat{\theta}_m) = I_1(m) + I_2(m) + I_3(m), \tag{4.4.1}$$

$$\text{with} \quad \begin{cases} I_1(m) &:= R(\widehat{\theta}_m) - R(\theta^*_m) \\ I_2(m) &:= \widehat{\gamma}_n(\theta^*_m) - \widehat{\gamma}_n(\widehat{\theta}_m) \\ I_3(m) &:= R(\theta^*_m) - \widehat{\gamma}_n(\theta^*_m) \end{cases} .$$

Next we provide a preliminary result about the asymptotic behavior of the terms $I_1(m)$ and $I_2(m)$. Then we obtain:

Lemma 4.1. *Under assumptions* **A0**-**A5**, *for any model $m \in \mathcal{M}$, there exists a probability distribution $U^*(m)$ such that*

1. $$n\,I_1(m) = n\left(R(\widehat{\theta}_m) - R(\theta^*_m)\right) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} U^*(m)$$
$$\text{and} \quad \mathbb{E}\big[n\,I_1(m)\big] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[U^*(m)] = \frac{1}{2}\,\mathit{Trace}\Big(\big(F(\theta^*_m)\big)^{-1}\,G(\theta^*_m)\Big). \tag{4.4.2}$$
2. $$n\,I_2(m) = n\left(\widehat{\gamma}_n(\theta^*_m) - \widehat{\gamma}_n(\widehat{\theta}_m)\right) \underset{n\to\infty}{\overset{\mathcal{D}}{\longrightarrow}} U^*(m)$$
$$\text{and} \quad \mathbb{E}\big[n\,I_2(m)\big] \underset{n\to\infty}{\longrightarrow} \frac{1}{2}\,\mathit{Trace}\Big(\big(F(\theta^*_m)\big)^{-1}\,G(\theta^*_m)\Big). \tag{4.4.3}$$

The proof of this lemma, as well as all the other proofs, can be found in Section 4.8. This result leads to our first main result devoted the ideal penalty defined in (4.2.10).

Proposition 4.2. *Under assumptions* **A0**-**A5** *and for any $m \in \mathcal{M}$, there exists a bounded sequence $(v^*_n)_{n\in\mathbb{N}^*}$ not depending on m satisfying*

$$\mathbb{E}\big[pen_{id}(m)\big] \underset{n\to\infty}{\sim} 2\left(\frac{1}{2n}\,\mathit{Trace}\Big(\big(F(\theta^*_m)\big)^{-1}\,G(\theta^*_m)\Big)\right) + \frac{v^*_n}{n}. \tag{4.4.4}$$

Note that the Slope Heuristic Procedure which allows to estimate a so called *minimal penalty* (see Arlot and Massart (2009)) consists in evaluating the slope of a linear regression of $\widehat{\gamma}_n(\widehat{\theta}_m)$ onto D_m for $m^* \subset m$ and this is equivalent to estimating the slope of $-\dfrac{1}{2n}\,\mathrm{Trace}\big(G(\theta^*_m)F(\theta^*_m)^{-1}\big)$ onto D_m. We will see that $\mathrm{Trace}\Big(\big(F(\theta^*_m)\big)^{-1}\,G(\theta^*_m)\Big)$ behaves as a linear function of D_m in many cases, which also gives legitimacy to this approach in the case of time series after having obtained it in the case of linear regression. The minimal penalty is then $-2\times$ the estimated slope and this finally corresponds to an estimation of $\mathbb{E}\big[\mathrm{pen}_{id}(m)\big]$.

Remark 4.2. *The trace of the matrix mentioned above is easily computable in some cases using the explicit forms of the matrices $F(\theta^*)$, $G(\theta^*)$ in Bardet and Wintenberger (2009). As showed in Section 4.6, this trace is proportionnal to the dimension of the model D_m in some cases but could be more complex functions of D_m.*

We can now state the main results of this paper.

Theorem 4.2. *Under assumptions* **A0**-**A5** *and if for any* $\varepsilon > 0$,

$$n\,\mathbb{P}\big(\mathrm{pen}(m) \geq \varepsilon\big) \underset{n\to\infty}{\longrightarrow} 0 \quad \textit{for any } m \in \mathcal{M}. \tag{4.4.5}$$

Then,

$$n\,\mathbb{P}\big(m^* \not\subset \widehat{m}_{\mathrm{pen}}\big) \underset{n\to\infty}{\longrightarrow} 0. \tag{4.4.6}$$

Theorem 4.2 says that if the penalty asymptotically decreases to 0 in probability, then the criterion $\widehat{C}_{\mathrm{pen}}$ does not select a misspecified model asymptotically.
Now, we can specify the convergence rate of pen to obtain an excess loss close to the minimal one over $\mathcal{M}$:

Theorem 4.3. *Under assumptions* **A0**-**A5**, *and if for any* $\varepsilon > 0$ *there exists* $K_\varepsilon > 0$ *such as*

$$\limsup_{n\to\infty} \max_{m\in\mathcal{M}} \mathbb{P}\Big(n\,\mathrm{pen}(m) \geq K_\varepsilon\Big) \leq \varepsilon. \tag{4.4.7}$$

Then for any $\varepsilon > 0$, *there exists* $M_\varepsilon > 0$ *and* $N_\varepsilon \in \mathbb{N}^*$ *such as for any* $n \geq N_\varepsilon$,

$$\mathbb{P}\Big(\ell(\widehat{\theta}_{\widehat{m}_{\mathrm{pen}}}, \theta^*) \leq \inf_{m\in\mathcal{M}} \big\{\ell(\widehat{\theta}_m, \theta^*)\big\} + \frac{M_\varepsilon}{n}\Big) \geq 1 - \varepsilon. \tag{4.4.8}$$

Remark 4.3. *Let notice that this asymptotic optimality is quite a bit different from the classical one about asymptotic efficiency, where both the cardinal of the collection* $\mathcal{M}$ *and the dimension of competitive models are allowed to depend on* n. *However, this is done in the framework where the parameter* θ^* *is infinite-dimensional (see for example Shibata (1980), Li (1987), Hsu et al. (2019b)).*

4.5 From a Bayesian model selection to a data-driven consistent model selection

4.5.1 Bayesian model selection

Another classical paradigm for model selection is the Bayesian one, leading typically to the BIC criterion (see Schwarz (1978)). In this approach, the construction of the model selection criterion is first done by assuming that the parameter vector θ^* is a random vector. Let recall the hierarchical prior sampling scheme in the Bayesian setting: given the finite family of models $\mathcal{M}$, a model m is drawn according to a prior distribution $(\pi_m)_{m\in\mathcal{M}}$ (generally a uniform distribution) and then, conditionally on m, θ is sampled according to some prior distribution $\mu_m(\theta)$.

The goal of this model selection procedure is to choose the most probable model after observing the trajectory $X := (X_1, \cdots, X_n)$, *i.e.*

$$\widehat{m}_B = \underset{m\in\mathcal{M}}{\mathrm{argmax}} \big\{\mathbb{P}\big(m \,|\, X\big)\big\}. \tag{4.5.1}$$

Using Bayes Formula, we can write $\mathbb{P}\big(m \,|\, X\big) = \frac{\pi_m\,\mathbb{P}\big(X \,|\, m\big)}{\mathbb{P}(X)}$. Moreover, we have:

$$\mathbb{P}\big(X \,|\, m\big) = \int_{\Theta_m} \mathbb{P}\big(X \,|\, \theta, m\big)\, d\mu_m(\theta).$$

In addition, since $\mathbb{P}(X)$ does not depend on m, and $\mathbb{P}\big(X \,|\, \theta, m\big)$ is the likelihood of X given $\theta \in \Theta_m$ and $m \in \mathcal{M}$, maximizing $\mathbb{P}\big(m \,|\, X\big)$ is equivalent to maximizing

$$S_n(m, X) := \log\big(\pi_m \, \mathbb{P}\big(X \,|\, m\big)\big) = \log\Big(\int_{\Theta_m} \pi_m \, \exp\big(L_n(\theta)\big) \, d\mu_m(\theta)\Big).$$

From now on, we will assume that $\pi_m = 1/|\mathcal{M}|$ for any $m \in \mathcal{M}$, a priori uniform distribution of the models in the family $\mathcal{M}$. We can also assume that there exists a non-negative Borel function $\theta \to b_m(\theta)$ such as $d\mu_m(\theta) = b_m(\theta)\, d\theta$. Then we have:

$$S_n(m, X) = -\log(|\mathcal{M}|) + \log\Big(\int_{\Theta_m} b_m(\theta) \, \exp\big(L_n(\theta)\big) \, d\theta\Big) \tag{4.5.2}$$

and by replacing L_n by the quasi version $\widehat{L}_n$, we introduce

$$\widehat{S}_n(m, X) = -\log(|\mathcal{M}|) + \log\Big(\int_{\Theta_m} b_m(\theta) \, \exp\big(L_n(\theta)\big) \, d\theta\Big). \tag{4.5.3}$$

Let us give an asymptotic expansion of the *a posteriori* probability $\widehat{S}_n(m, X)$ in order to derive a BIC type criterion that is coherent with our framework where the observed trajectory is that of a causal affine process. This could be obtained from a Laplace approximation leading to the following theorem:

Theorem 4.4. *Under assumptions* **A0, A1, A2, A3, A5** *and if for any $x \in \mathbb{R}^\infty$, the functions $\theta \to M_\theta$ and $\theta \to f_\theta$ are $\mathcal{C}^6(\Theta)$ functions satisfying $\mathbf{A}(\partial^k_{\theta^k} f_\theta, \Theta)$ and $\mathbf{A}(\partial^k_{\theta^k} M_\theta, \Theta)$ for any $0 \le k \le 2$. Then,*

$$\begin{aligned}\widehat{S}_n(m, X) = \widehat{L}_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m| + \log\big(b_m(\widehat{\theta}_m)\big) \\ + \frac{\log(2\pi)}{2}\,|m| - \frac{1}{2}\,\log\Big[\det\big(-\widehat{F}_n(m)\big)\Big] - \log(|\mathcal{M}|) + O(n^{-1}) \quad a.s.\end{aligned} \tag{4.5.4}$$

where $\widehat{F}_n(m) := \Big(\partial^2_{\theta_i\theta_j} \widehat{L}_n(\widehat{\theta}_m)\Big)_{i,j\in m}$

In the above equation, it is clear that $-2\,\widehat{S}_n(m, X) \simeq -2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m|$ a.s. This gives legitimacy to the usual BIC criterion within the framework of causal affine processes since:

$$\widehat{m}_{BIC} = \text{argmin}_{m\in\mathcal{M}} \Big\{ -2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m| \Big\},$$

and we see that $\widehat{m}_{BIC}$ maximizes the main terms of $\widehat{S}_n(m, X)$.
From the relation (4.5.4), considering certain second order terms of the asymptotic expansion of $\widehat{S}_n(m, X)$, we also obtain the Kashyap criterion (see Kashyap (1982), Sclove (1987), Bozdogan (1987)), defined for all $m \in \mathcal{M}$ by

$$\begin{aligned}\widehat{KC}(m) := -2\,\widehat{L}_n(\widehat{\theta}_m) + \log(n)\,|m| + \log\Big[\det\big(-\widehat{F}_n(m)\big)\Big] \\ \text{and} \quad \widehat{m}_{KC} = \text{argmin}_{m\in\mathcal{M}} \big\{\widehat{KC}(m)\big\}.\end{aligned} \tag{4.5.5}$$

Therefore the term $\log\big[\det\big(\widehat{F}_n(m)\big)\big)$ is added to the usual BIC criterion. Several example of computations of this term, generally equal to $c\,|m|$ but not always, are provided in the forthcoming Section 4.6. It is clear that $\widehat{m}_{KC}$ can be more interesting that $\widehat{m}_{BIC}$ in terms of consistency olny for non asymptotic framework (typically for n of the order of a hundred

or several hundred). Note also that the data-driven criteria $\widehat{KC}$ that is "optimal" in the sense of the a posteriori probability (see Kashyap Kashyap (1982)) is also asymptotically consistent under the assumption **A5**, see Bardet *et al.* Bardet et al. (2020b).
However this choice of second order terms of the asymptotic expansion of $\widehat{S}_n(m, X)$ is somewhere arbitrary. A criterion taking account of all the second order terms could also be defined. For this, we could define a uniform distribution b_m on a compact set included in Θ_m. As a consequence, using condition (4.3.1) of Assumption **A0**, there always exists $0 < C_1 \leq C_2$ such as $\frac{C_1}{m} \leq b_m(\theta_m) \leq \frac{C_2}{m}$. As a consequence, we could define a new data-driven consistent criterion, called $\widehat{KC}'$, such as for any $m \in \mathcal{M}$

$$\widehat{KC}'(m) := -2\,\widehat{L}_n(\widehat{\theta}_m) + \big(\log(n) - \log(2\pi)\big)\,|m| + \log\Big[\det\big(-\widehat{F}_n(m)\big)\Big] + 2\log\big(|m|\big)$$
$$\text{and} \quad \widehat{m}_{KC'} = \operatorname{argmin}_{m\in\mathcal{M}}\big\{\widehat{KC}'(m)\big\}. \quad (4.5.6)$$

Remark 4.4. *We also know that under assumptions* **A0, A1, A2, A3, A5**, $\widehat{F}_n(m) \xrightarrow[n\to+\infty]{a.s.} F(\theta^*_m)$. *Therefore* $\log\big(\det\big(F(\theta^*_m)\big)\big)$ *can also replace* $\log\big(\det\big(\widehat{F}_n(m)\big)\big)$ *in the expression of* $\widehat{KC}'(m)$.

Corollary 4.2. *Under assumptions* **A0, A1, A2, A3** *and* **A5**, *from Bardet et al. (2020b),* $\widehat{m}_{BIC}$, $\widehat{m}_{KC}$ *and* $\widehat{m}_{KC'}$ *are consistent criteria, i.e.*

$$\widehat{m}_{BIC} \xrightarrow[n\to\infty]{\mathcal{P}} m^*, \quad \widehat{m}_{KC} \xrightarrow[n\to\infty]{\mathcal{P}} m^* \quad \text{and} \quad \widehat{m}_{KC'} \xrightarrow[n\to\infty]{\mathcal{P}} m^*.$$

4.6 Computations of the ideal penalties

4.6.1 Computations of Trace$\Big(\big(F(\theta^*_m)\big)^{-1} G(\theta^*_m)\Big)$

From Bardet and Wintenberger (2009), with $\mu_4 = \mathbb{E}[\xi_0^4]$, f^0_θ and H^0_θ defined in (1.2.3), we have for $m^* \subset m$ and $i, j \in m$:

$$\begin{aligned}
\big(G(\theta^*_m)\big)_{i,j} &= \mathbb{E}\Big[\frac{\partial_{\theta_i} f^0_{\theta^*_m}\,\partial_{\theta_j} f^0_{\theta^*_m}}{H^0_{\theta^*_m}} + \frac{(\mu_4-1)}{4}\,\frac{\partial_{\theta_i} H^0_{\theta^*_m}\,\partial_{\theta_j} H^0_{\theta^*_m}}{(H^0_{\theta^*_m})^2}\Big] \\
\big(F(\theta^*_m)\big)_{i,j} &= -\mathbb{E}\Big[\frac{\partial_{\theta_i} f^0_{\theta^*_m}\,\partial_{\theta_j} f^0_{\theta^*_m}}{H^0_{\theta^*_m}} + \frac{1}{2}\,\frac{\partial_{\theta_i} H^0_{\theta^*_m}\,\partial_{\theta_j} H^0_{\theta^*_m}}{(H^0_{\theta^*_m})^2}\Big], \qquad (4.6.1)
\end{aligned}$$

Here there are 3 frameworks where Trace$\Big(\big(F(\theta^*_m)\big)^{-1} G(\theta^*_m)\Big)$ can be computed for $m^* \subset m$:

1/ A first and well-known case is the Gaussian case. Indeed, when (ξ_t) is a Gaussian white noise, then $\mu_4 = 3$ and then from (4.6.1), for any $i, j \in m$,

$$\big(G(\theta^*_m)\big)_{i,j} = -\big(F(\theta^*_m)\big)_{i,j} \implies -2\,\text{Trace}\big(\big(F(\theta^*_m)\big)^{-1} G(\theta^*_m)\big) = 2\,\text{Trace}\big(I_{|m|}\big) = 2\,|m|,$$

with I_ℓ the identity matrix of size $\ell \in \mathbb{N}^*$. As a consequence, in the Gaussian framework, for $m^* \subset m$, the expectation of the ideal penalty is exactly the classical Akaike Criterion (AIC).

2/ A frequent case is when the parameter θ identifying an affine causal model $X_t = M^t_\theta\,\xi_t + f^t_\theta$ can be decomposed as $\theta = (\theta_1, \theta_2)'$ with $f^t_\theta = \widetilde{f}^t_{\theta_1}$ and $M^t_\theta = \widetilde{M}^t_{\theta_2}$. Let p_1, p_2

such that $p_1 = |\theta_1|$, $p_2 = |\theta_2|$ and $|m| = p_1 + p_2$.
In such a case, from (4.6.1), it is clear that all the terms $F(\theta_m^*)_{i,j}$ and $G(\theta_m^*)_{i,j}$ are equals to zero for $i = 1, \dots, p_1$ and $j = 1, \cdots, p_2$ implying

$$F(\theta_m^*) = -\begin{pmatrix} A_{1,p_1} & O_{p_1,p_2} \\ O_{p_2,p_1} & B_{p_1+1,p_1+p_2} \end{pmatrix} \text{ and } \quad G(\theta_m^*) = \begin{pmatrix} A_{1,p_1} & O_{p_1,p_2} \\ O_{p_2,p_1} & \frac{(\mu_4-1)}{2} B_{p_1+1,p_1+p_2} \end{pmatrix}$$

where O is the null matrix and from the expressions of matrix $G(\theta_m^*)$ and $F(\theta_m^*)$ in (4.6.1), $A_{1,p_1} = \Big(\mathbb{E}\Big[\frac{\partial_{\theta_i} f^0_{\theta_m^*} \, \partial_{\theta_j} f^0_{\theta_m^*}}{H^0_{\theta_m^*}}\Big]\Big)_{1\le i,j\le p_1}$ and $\quad B_{p_1+1,p_1+p_2} = \Big(\frac{1}{2}\,\mathbb{E}\Big[\frac{\partial_{\theta_i} H^0_{\theta_m^*} \, \partial_{\theta_j} H^0_{\theta_m^*}}{(H^0_{\theta_m^*})^2}\Big]\Big)_{p_1+1\le i,j\le p_1+p_2}$.
As a consequence,

$$\begin{aligned} G(\theta_m^*)\, F(\theta_m^*)^{-1} &= -\mathrm{Diag}\Big(A_{1,p_1}, \frac{(\mu_4-1)}{2} B_{p_1+1,p_1+p_2}\Big) \times \mathrm{Diag}\big(A^{-1}_{1,p_1}, B^{-1}_{p_1+1,p_1+p_2}\big) \\ &= -\mathrm{Diag}\big(I_{p_1}, \frac{(\mu_4-1)}{2} I_{p_2}\big) \end{aligned}$$

and we obtain

$$-2\,\mathrm{Trace}\Big(\big(F(\theta_m^*)\big)^{-1} G(\theta_m^*)\Big) = 2\,p_1 + (\mu_4 - 1)\,p_2. \tag{4.6.2}$$

This setting includes many classical times series:

- For ARMA(p,q) processes, we have $X_t = f^t_\theta + \sigma\,\xi_t$ since $X_t + a_1\,X_{t-1} + \cdots + a_p\,X_{t-p} = \sigma\,(\xi_t + b_1\,\xi_{t-1} + \cdots + b_q\,\xi_{t-q})$ for all $t \in \mathbb{Z}$. Then $\theta_1 = (a_1, \dots, a_p, b_1, \dots, b_q)$ and $\theta_2 = \sigma$. The penalty term is slightly different according to σ is known or not:

  (a) if σ is known, then $\theta = \theta_1$ and $G(\theta^*) = -F(\theta^*)$, so that we recover exactly the AIC penalty term:

  $$-2\,\mathrm{Trace}\big(G(\theta_m^*)F(\theta_m^*)^{-1}\big) = 2\,|m| = 2\,(p+q);$$

  (b) Otherwise, $\theta = (\theta_1, \sigma)$ and simple computations lead to

  $$F(\theta^*) = \begin{pmatrix} \big(F(\theta^*)\big)_{1\le i,j\le |m|-1} & 0 \\ 0 & -\frac{1}{2\,\sigma^4} \end{pmatrix} \text{ and } G(\theta^*) = \begin{pmatrix} \big(G(\theta^*)\big)_{1\le i,j\le |m|-1} & 0 \\ 0 & \frac{(\mu_4-1)}{4\,\sigma^4} \end{pmatrix}$$

  where $\big(G(\theta^*)\big)_{1\le i,j\le |m|-1} = -\big(F(\theta^*)\big)_{1\le i,j\le |m|-1}$.
  Thus, we obtain $G(\theta^*)F(\theta^*)^{-1} = -\begin{pmatrix} I_{1\le i,j\le |m|-1} & 0 \\ 0 & \frac{\mu_4-1}{2} \end{pmatrix}$ and therefore, with $|m| = p+q+1$ in this case,

  $$-2\,\mathrm{Trace}\big(G(\theta_m^*)F(\theta_m^*)^{-1}\big) = 2\,|m| + (\mu_4 - 3) = 2(p+q) + (\mu_4 - 1),$$

  and therefore once again the expectation of the ideal penalty leads to the AIC model selection.

- For GARCH(p,q) processes (see Francq and Zakoian (2010)), we have $f_\theta = 0$ and $X_t = M^t_\theta\,\xi_t$ since for any $t \in \mathbb{Z}$,

  $$\begin{cases} X_t &= \sigma_t\,\xi_t \\ \sigma_t^2 &= \omega_0 + a_1\,X^2_{t-1} + \cdots + a_p\,X^2_{t-p} + b_1\,\sigma^2_{t-1} + \cdots + b_q\,\sigma^2_{t-q} \end{cases}.$$

  Denote $\theta = \theta_2 = (\omega_0, a_1, \dots, a_p, b_1, \dots, b_q)$.
  Then we have $A_{p_1} = 0$ and therefore $G(\theta^*) = -\frac{(\mu_4-1)}{2}\,F(\theta^*)$. As a result:

  $$-2\,\mathrm{Trace}\big(G(\theta_m^*)F(\theta_m^*)^{-1}\big) = (\mu_4 - 1)\,|m| = (\mu_4 - 1)\,(p+q+1).$$

- For APARCH(δ, p, q) processes (see Ding et al. (1993)), we also have $f_\theta = 0$ and $X_t = M_\theta^t \, \xi_t$ since for any $t \in \mathbb{Z}$,

$$\left\{ \begin{array}{rcl} X_t & = & \sigma_t \, \xi_t \\ \sigma_t^\delta & = & \omega_0 + a_1 \, (X_{t-1} - \gamma_1 |X_{t-1}|)^\delta + \cdots + a_p \, (X_{t-p} - \gamma_p |X_{t-p}|)^\delta \\ & & \qquad + b_1 \, \sigma_{t-1}^\delta + \cdots + b_q \, \sigma_{t-q}^\delta \end{array} \right. .$$

  For such a process, $\theta = \theta_2 = (\omega_0, a_1, \ldots, a_p, \gamma_1, \ldots, \gamma_p, b_1, \ldots, b_q)$ when we assume that δ is known, and, *mutatis mutandis*, the result is the same than for GARCH processes:

$$-2 \, \mathrm{Trace}\big(G(\theta_m^*) F(\theta_m^*)^{-1}\big) = (\mu_4 - 1) \, |m| = (\mu_4 - 1) \, (2p + q + 1).$$

3/ Otherwise, the computations are no longer easy. Let us see the example of the family of $AR(1) - ARCH(p)$ processes. Then for any $t \in \mathbb{Z}$ we have $X_t = \phi X_{t-1} + Z_t$ where $Z_t = \xi_t \left(\alpha_0 + \alpha_1 Z_{t-1}^2 + \cdots + \alpha_p Z_{t-p}^2\right)^{1/2}$. As a consequence, with $\theta = (\phi, \alpha_0, \ldots, \alpha_p)'$, we obtain for any $t \in \mathbb{Z}$,

$$X_t = f_\theta(X_{t-1}) + M_\theta(X_{t-1}, \ldots, X_{t-p-1}) \, \xi_t$$
$$\text{with} \quad \left\{ \begin{array}{lcl} f_\theta(X_{t-1}) & = & \phi \, X_{t-1} \\ M_\theta(X_{t-1}, \ldots, X_{t-p}) & = & \left(\alpha_0 + \sum_{i=1}^p \alpha_i (X_{t-i} - \phi X_{t-i-1})^2\right)^{1/2} \end{array} \right. .$$

Thus the parameter ϕ is present in f_θ as well as in M_θ. From (4.6.1), and with the notations of 1/, we obtain:

$$F(\theta_m^*) = -\begin{pmatrix} A_{1,1} & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix} - B_{1,p+2}$$
$$\text{and} \quad G(\theta_m^*) = \begin{pmatrix} A_{1,1} & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix} + \frac{(\mu_4 - 1)}{2} \, B_{1,p+2}.$$

As a consequence,

$$G(\theta_m^*) = -\frac{(\mu_4 - 1)}{2} \, F(\theta_m^*) + \frac{(\mu_4 - 3)}{2} \begin{pmatrix} A_{1,1} & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix}.$$

Thus, with $|m| = p + 2$,

$$G(\theta_m^*) \, F^{-1}(\theta_m^*) = -\frac{(\mu_4 - 1)}{2} \, I_{|m|} + \frac{(\mu_4 - 3)}{2} \begin{pmatrix} A_{1,1} & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix} F^{-1}(\theta_m^*).$$

Whatever the matrix $F^{-1}(\theta_m^*)$, we have $\begin{pmatrix} A_{1,1} & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix} F^{-1}(\theta_m^*) = \begin{pmatrix} c(\theta_m^*) & O_{1,p+1} \\ O_{p+1,1} & O_{p+1,p+1} \end{pmatrix}$ with $c(\theta_m^*) = c(\theta^*) \in \mathbb{R}$ since $m^* \subset m$. Then for all $m^* \subset m$,

$$-2 \, \mathrm{Trace}\big(G(\theta_m^*) F(\theta_m^*)^{-1}\big) = -2 \, c(\theta^*) + (\mu_4 - 1) \, |m|,$$

where $-2 \, c(\theta^*)$ does not depend on m.

4.7 Numerical Studies

This section aims to investigate how well are our new model selection criteria based on the theoritical results obtained in both Section 4.4 and Section 4.5.
To do that, three Data Generating Process (DGP) have been considered:

$$\begin{aligned}
\text{DGP I} \quad \text{AR}(2) \qquad & X_t = 0.4\, X_{t-1} + 0.4\, X_{t-2} + \xi_t, \\
\text{DGP II} \quad \text{ARMA}(1,1) \qquad & X_t - 0.5\, X_{t-1} = \xi_t + 0.6\, \xi_{t-1}, \\
\text{DGP III} \quad \text{GARCH}(1,1) \qquad & X_t = \sigma_t\, \xi_t \quad \text{with } \sigma_t^2 = \left(1 + 0.35 X_{t-1}^2 + 0.4 \sigma_{t-1}^2\right)^{1/2},
\end{aligned}$$

where $(\xi_t)_t$ is a Gaussian white noise with variance unity.

In order to identify possible discrepancies between asymptotically expected results and those obtained at finite distance, we consider n belongs to 200, 500, 1000, 2000. We will compare the performance of the AIC, BIC and KC'.

For the efficiency property (Theorem 1.2), we will be interested in the difference

$$ME := n\left(\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) - \inf_{m \in \mathcal{M}} \left\{ \ell(\widehat{\theta}_m, \theta^*) \right\}\right)$$

and the idea would be to see the decay rate of this residual term.

The considered family of competitive models is the same for the three DGP

$$\mathcal{M} = \left\{ \text{ARMA}(p,q) \ \text{ or } \ \text{GARCH}(p,q) \quad \text{with} \quad 0 \le p,\ q \le 6 \right\}.$$

The empirical estimates of the ratio ME are obtained based on 500 replications. Here is how we empirically estimated ME.
Given a DGP and a list of candidate models, the empirical estimates of $\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)$ have been obtained in the following way:

1. For any model m, the observations $(X_1, \ldots, X_n)$ from DGP have been used to estimated the QMLE $\widehat{\theta}_m$ and then we considered as an estimate of $R(\widehat{\theta}_m)$ the quantity $\widehat{\gamma}_n(\widehat{\theta}_m)$ which is the empirical risk of the estimator $\widehat{\theta}_m$ but computed over observations $(Y_1, \ldots, Y_N)$ (with $N = 10^6$) arisen from the DGP and independent to $(X_1, \ldots, X_n)$.

2. Selection of $\widehat{m}$ according to (4.2.9) and then we retrieve the estimation of $R(\widehat{\theta}_{\widehat{m}})$ in the list of estimates obtained in the first point;

3. Then take the difference of the results obtained in 1 and 2. The experiment was repeated 500 times and the average of these differences times n is our estimate of ME.

For the consistency property, the Table 4.1 shows the frequency of selecting respectively the true model versus a model other than the true one (called here "wrong").

Table 4.1: Percentage of selected order based on 500 replications depending on sample's length for DGP I, DGP II and DGP III respectively.

Sample length	200			500			1000			2000		
	AIC	BIC	KC'	AIC	BIC	KC'	AIC	BIC	KC'	AIC	BIC	KC'
True	17.2	36.2	35.6	30.4	73.2	78.2	36.4	87.4	92.2	32.4	96.2	98.4
Wrong	82.8	63.8	64.4	69.6	26.8	21.8	63.6	13.6	7.8	67.6	3.8	1.6
True	27.8	80.8	92.0	30.6	88.4	96.6	31.0	89.1	97.5	33.3	95.2	99.9
Wrong	72.2	19.2	8.0	69.7	11.6	3.4	69.0	10.9	2.5	66.7	4.8	0.1
True	0.4	10.8	14.8	1.4	32.2	55.8	1.0	54.8	82.0	2.0	75.8	93.8
Wrong	99.6	89.2	85.2	98.6	67.8	44.2	99.0	45.2	18.0	98.0	24.2	6.2

From these results, it follows that KC' outperforms BIC when dealing both with small and larges samples. These results confirm that it is important to consider also the neglected terms in the derivation of the BIC criterion.
Moreover, the percentage of selecting the true model using both BIC and KC' approaches 100 with increasing n. This is consistent with our asymptotic result (Corollary 4.2). Hence, KC' is more robust to the sample size and thus improves BIC.

Table 4.2: ME estimates based on 500 replications depending on sample's length.

Sample length	200			500			1000			2000		
	AIC	BIC	KC'	AIC	BIC	KC'	AIC	BIC	KC'	AIC	BIC	KC'
DGP I	4.91	2.59	5.35	3.46	1.11	1.18	3.08	0.98	0.75	3.05	0.38	0.29
DGP II	3.66	0.87	0.54	3.37	0.42	0.11	2.62	0.15	0.05	2.5	0.10	0.04
DGP III	2.39	4.63	13.16	2.53	4.08	9.54	2.69	2.96	2.52	3.21	2.06	0.76

In view of the results of Table 4.2, we notice a decrease of the residual term $\widehat{ME}$ for all the criteria. This decrease is much faster and tends towards 0 for the consistent criteria while it is much less for the AIC. This allows us to conclude that Theorem 4.3 is verified by AIC criterion. Moreover, Theorem 4.3 is also verified by both criteria BIC and KC' despite the fact that these criteria do not satisfy the conditions of the theorem. One might think that the validity of (4.4.8) is due to the fact that since $m^* \in \mathcal{M}$, $\inf_{m\in\mathcal{M}}\{\ell(\widehat{\theta}_m, \theta^*)\}$ is reached for $m = m^*$ in most cases and to the consistency of the BIC and KC' criteria. Thus, one could say that in configurations where the true model is part of the candidate models, the consistent criteria are also efficient.

4.8 Proofs

4.8.1 Proofs of Section 4.3

The asymptotic normality of $\left(\frac{1}{n}\left(\partial_{\theta_i} L_n(\theta^*_m)\right)_{i\in m}\right.$ was established in Bardet and Wintenberger (2009) and Bardet et al. (2020b) when $m^* \subset m$ using a central limit theorem for stationary martingale difference. Here we extend this result to any $m \in \mathcal{M}$:

Proposition 4.3. *Under Assumption A0-A5, for any $\theta \in \Theta$, we have*

$$\sqrt{n}\Big(\frac{1}{n}\,\partial_\theta L_n(\theta) + \frac{1}{2}\,\mathbb{E}\big[\partial_\theta \gamma(\theta, X_0)\big]\Big) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\big(0\,,\,\Sigma(\theta)\big)$$
$$\text{with}\quad \Sigma(\theta) := \frac{1}{4}\Big(\sum_{t\in\mathbb{Z}} Cov\left(\partial_{\theta_i}\gamma(\theta, X_0)\,,\,\partial_{\theta_j}\gamma(\theta, X_t)\right)\Big)_{1\le i,j\le d}. \quad (4.8.1)$$

The main tool we use here for establishing Theorem 4.3 is the notion of τ-dependence for stationary time series. More precisely, the τ-dependence coefficients, which are a version of the coupling coefficients introduced in Dedecker and Prieur (2004) and used for stationary infinite memory chains. The reader is deferred to the lecture notes Dedecker et al. (2007) for complements and details on coupling, based on the Wasserstein distance between probabilities defined as below. Its stationary version is: Let $(\Omega, \mathcal{C}, \mathbb{P})$ be a probability space, $\mathcal{M}$ a σ-subalgebra of $\mathcal{C}$ and Z a random variable with values in E. Assume that $\|Z\|_p < \infty$ and define the coefficient $\tau^{(p)}$ as

$$\tau^{(p)}(\mathcal{M}, Z) = \Big\| \sup_{f\in\Lambda_1(E)} \Big\{ \Big| \int f(x)\mathbb{P}_{Z|\mathcal{M}}(dx) - \int f(x)\mathbb{P}_Z(dx)\Big| \Big\}\Big\|_p.$$

Using the definition of τ, the dependence between the past of the sequence $(Z_t)_{t\in\mathbb{Z}}$ and its future k-tuples may be assessed: consider the norm $\|x-y\| = \|x_1-y_1\| + \cdots + \|x_k - y_k\|$ on E^k, set $\mathcal{M}_p = \sigma(Z_t, t \le p)$ and define

$$\tau_Z^{(p)}(s) \;=\; \sup_{k>0}\Big\{ \max_{1\le l\le k} \frac{1}{l} \sup\Big\{\tau^{(p)}(\mathcal{M}_p, (Z_{j_1}, \ldots, Z_{j_l})) \text{ with } p + s \le j_1 < \cdots < j_l\Big\}\Big\}.$$

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

Lemma 4.2. *Under Assumption A0, then for $p \le r$ and $b_k^{(p)} = \|\xi_0\|_p\,\alpha_k(M_\theta, \Theta) + \alpha_k(f_\theta, \Theta)$ for any $j \in \mathbb{N}^*$,*

$$\tau_X^{(p)}(s) \le C\,\lambda_s \quad \text{with} \quad \lambda_s = \inf_{1\le r\le s}\Big\{\Big(\sum_{k=1}^{\infty} b_k^{(p)}\Big)^{s/r} + \sum_{t=r+1}^{\infty} b_t^{(p)}\Big\} \quad \text{for } s \ge 1. \qquad (4.8.2)$$

Proof of Lemma 4.2. This Lemma can be directly deduced from Proposition 3.1 of Doukhan and Wintenberger (2008) where $F(x, \xi_0) = M_\theta(x)\,\xi_0 + f_\theta(x)$ for any $x \in \mathbb{R}^\infty$ and therefore

$$\big\|F(x,\xi_0) - F(y,\xi_0)\big\|_p \le \|\xi_0\|_p\,\big|M_\theta(x) - M_\theta(y)\big| + \big|f_\theta(x) - f_\theta(y)\big|$$

inducing $\big\|F(x,\xi_0) - F(y,\xi_0)\big\|_p \le \sum_{k=1}^{\infty} b_k^{(p)}$ with $b_k^{(p)} = \|\xi_0\|_p\,\alpha_k(M_\theta,\Theta) + \alpha_k(f_\theta,\Theta)$. □

Remark 4.5. *Using Assumption A0 and A5, we deduce that $b_t^{(p)} = O\big(t^{-\delta}\big)$ with $\delta > 7/2$, and therefore $\tau_X^{(p)}(s) \le \lambda_s = O\big((s^{-1}\log s)^{\delta-1}\big)$.*

Now, under Assumption A0, since X is a causal time series, define for any $j = 1, \ldots, d$ and $\theta \in \Theta$,

$$\phi_\theta^{(j)}\big((X_{t-k})_{k\ge 0}\big) := \partial_{\theta_j}\gamma(\theta, X_t) = -2\,\partial_{\theta_j} M_\theta^t\,\frac{(X_t - f_\theta^t)^2}{\big(M_\theta^t\big)^3} - 2\,\partial_{\theta_j} f_\theta^t\,\frac{X_t - f_\theta^t}{\big(M_\theta^t\big)^2} + 2\,\frac{\partial_{\theta_j} M_\theta^t}{M_\theta^t}.$$

Then we have:

Lemma 4.3. *Under Assumption A0-A5, for any $j = 1, \dots, d$, for any $\theta \in \Theta$, the sequence $\big(\phi_\theta^{(j)}\big((X_{t-k})_{k\geq 0}\big)\big)_{t\in\mathbb{Z}}$ is a causal stationary sequence that is $\tau^{(p)}_{\phi_\theta^{(j)}}$-weakly dependent where its coefficients $\tau^{(1)}_{\phi_\theta^{(j)}}(s)$ satisfies:*

$$\tau^{(1)}_{\phi_\theta^{(j)}}(s) \leq C\Big(\sum_{\ell=1}^{s}\big(\alpha_\ell(f_\theta,\Theta)+\alpha_\ell(M_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}M_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}f_\theta,\Theta)\big)\,\lambda_{s+1-\ell} + \sum_{\ell=s+1}^{\infty}\big(\alpha_\ell(f_\theta,\Theta)+\alpha_\ell(M_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}M_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}f_\theta,\Theta)\big)\Big), \quad (4.8.3)$$

for any $s \geq 0$ where (λ_s) is defined in (4.8.2).

Proof of Lemma 4.3. In the proof of Proposition 4.1 of Bardet et al. (2020a), it has been proven for $U = (U_i)_{i\geq 1}$ and $V = (V_i)_{i\geq 1}$ such as $\sup_{i\geq 1}\big\{\|U_i\|_4 \vee \|V_i\|_4\big\} < \infty$ that there exists $C > 0$ satisfyng

$$\mathbb{E}\big[\sup_{\theta\in\Theta}\big\|\phi_\theta^{(j)}(U)-\phi_\theta^{(j)}(V)\big\|\,\big] \leq C\Big(\|U_1-V_1\|_4 + \sum_{i=2}^{\infty}\big(\alpha_i(f_\theta,\Theta)+\alpha_i(M_\theta,\Theta)+\alpha_i(\partial_{\theta_j}f_\theta,\Theta)+\alpha_i(\partial_{\theta_j}M_\theta,\Theta)\big)\|U_i-V_i\|_4\Big). \quad (4.8.4)$$

Using coupling techniques, if $(\widetilde{\xi}_t)_{t\in\mathbb{Z}}$ is an independent replication of $(\xi_t)_{t\in\mathbb{Z}}$, define also $(\widetilde{X}_t)_{t\in\mathbb{Z}}$ satisfying Assumption with $(\widetilde{\xi}_t)_{t\in\mathbb{Z}}$ instead of $(\xi_t)_{t\in\mathbb{Z}}$ and $\big(\phi_\theta^{(j)}\big((\widetilde{X}_{t-k})_{k\geq 0}\big)\big)_{t\in\mathbb{Z}}$. Then for $s \geq 0$, using (4.8.4),

$$\begin{aligned}
\tau^{(1)}_{\phi_\theta^{(j)}}(s) &\leq \big\|\phi_\theta^{(j)}\big((X_{s-k})_{k\geq 0}\big)-\phi_\theta^{(j)}\big((\widetilde{X}_{s-k})_{k\geq 0}\big)\big\|_1 \\
&\leq C\Big(\|X_1-\widetilde{X}_1\|_4 + \sum_{i=2}^{\infty}\big(\alpha_i(f_\theta,\Theta)+\alpha_i(M_\theta,\Theta)+\alpha_i(\partial_{\theta_j}f_\theta,\Theta)+\alpha_i(\partial_{\theta_j}M_\theta,\Theta)\big)\|X_i-\widetilde{X}_i\|_4\Big) \\
&\leq C\sum_{\ell=1}^{\infty}\big(\alpha_\ell(f_\theta,\Theta)+\alpha_\ell(M_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}f_\theta,\Theta)+\alpha_\ell(\partial_{\theta_j}M_\theta,\Theta)\big)\,\lambda_{s+1-\ell},
\end{aligned}$$

that implies (4.8.3). □

Remark 4.6. *Under Assumption A0 and A5, and therefore with $\lambda_s = O\big(s^{1-\delta}\log s\big)$ with $\delta > 7/2$, we also deduce that $\tau^{(1)}_{\phi_\theta^{(j)}}(s) = O\big(s^{1-\delta}\log s\big)$.*

Proof of Proposition 4.3. If Z is a τ_Z-dependent centered stationary time series satisfying $\mathbb{E}[|Z_0|^\kappa] < \infty$ with $\kappa > 2$, and $\sum_{s=1}^{\infty} s^{1/(\kappa-2)}\tau_Z(s) < \infty$, we deduce from Lemma 2, point 2. of Dedecker and Doukhan (2003) that condition $\mathrm{D}(2,\theta/2,X)$ is satisfied as θ-weakly dependent coefficients are smaller than τ-weakly dependent coefficients, see (2.2.13) p.16 of Dedecker et al. (2007), and $0 < \sum_{t\in\mathbb{Z}}\big|\mathbb{E}[Z_0Z_t]\big| < \infty$ from Proposition 2 of Dedecker and Doukhan (2003). Then,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\Big(0\,,\sum_{t\in\mathbb{Z}}\mathbb{E}[Z_0Z_t]\Big).$$

We can apply this central limit theorem to

$$Z_t := \sum_{j=1}^{d} c_j \big(\phi_\theta^{(j)}\big((X_{t-k})_{k\geq 0}\big) - \mathbb{E}\big[\phi_\theta^{(j)}\big((X_{t-k})_{k\geq 0}\big)\big]\big) \quad \text{with } (c_j)_{1\leq j\leq d} \in \mathbb{R}^d.$$

Indeed, using Lemma 4.3, we easily obtain for $s \geq 0$

$$\tau_Z(s) \leq C\Big(\sum_{j=1}^{d} |c_j|\Big) \sum_{\ell=1}^{\infty} \big(\alpha_\ell(f_\theta,\Theta) + \alpha_\ell(M_\theta,\Theta) + \alpha_\ell(\partial_{\theta_j} f_\theta,\Theta) + \alpha_\ell(\partial_{\theta_j} M_\theta,\Theta)\big)\, \lambda_{s+1-\ell},$$

and therefore under Assumption A0 and A5, $\tau_Z(s) = O\big(s^{1-\delta}\log s\big)$. Moreover, using Lemma 4.6, we deduce $\mathbb{E}\big[|Z_0|^{8/3}\big] < \infty$. Then with $\kappa = 8/3$, $\sum_{s=1}^{\infty} s^{1/(\kappa-2)}\tau_Z(s) = \sum_{s=1}^{\infty} s^{3/2}\,\tau_Z(s) < \infty$ is satisfied since $\delta > 7/2$.
Therefore, we deduce for any $\theta \in \Theta$,

$$\sqrt{n}\sum_{j=1}^{d} c_j \Big(\frac{1}{n}\,\partial_{\theta_j} L_n(\theta) + \frac{1}{2}\,\mathbb{E}\big[\partial_{\theta_j}\gamma(\theta, X_0)\big]\Big)$$

$$\xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\Big(0\,,\, \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d} c_i\, c_j \sum_{t\in\mathbb{Z}} \mathrm{Cov}\left(\partial_{\theta_i}\gamma(\theta, X_0)\,,\, \partial_{\theta_j}\gamma(\theta, X_t)\right)\Big),$$

which implies the multidimensional central limit theorem (4.8.1). □

Proof of Corollary 4.1. Firstly, it was already established in Bardet et al. (2020b) that if $m^* \subset m$ then $\big(\partial_{\theta_i}\gamma(\theta_m, X_t)\big)_{t\in\mathbb{Z}}$ is a stationary martingale increments with respect to $\mathcal{F}_t = \sigma\big((X_{t-k})_{k\in\mathbb{N}}\big)$. As a consequence $\mathrm{Cov}\left(\partial_{\theta_i}\gamma(\theta, X_0)\,,\, \partial_{\theta_j}\gamma(\theta, X_t)\right) = 0$ if $t \neq 0$.
Secondly, for all $m \in \mathcal{M}$, from the definition of θ_m^* as a local minimum of R on Θ_m, and from Assumption A0-A5, then $\partial_{\theta_j} R(\theta_m^*) = \mathbb{E}\big[\partial_{\theta_j}\gamma(\theta_m^*, X_0)\big] = 0$ for all $j \in m$. □

Proof of Theorem 4.1. We use here the standard proof allowing to show the asymptotic normality of the QMLE and already used in Bardet and Wintenberger (2009).
Firstly, it was established in Bardet et al. (2020b) that $\widehat{\theta}_m \xrightarrow[n\to+\infty]{a.s.} \theta_m^*$.
Secondly, a Taylor-Lagrange expansion is applied to $\big(\partial_{\theta_j} L_n(\widehat{\theta}_m)\big)_{j\in m}$ around θ_m^*:

$$\frac{1}{\sqrt{n}}\left(\partial_{\theta_j} L_n(\widehat{\theta}_m)\right)_{j\in m} = \frac{1}{\sqrt{n}}\left(\partial_{\theta_j} L_n(\theta_m^*)\right)_{j\in m}$$

$$+ \Big(\frac{1}{n}\,\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_m)\Big)_{i,j\in m} \times \sqrt{n}\left((\widehat{\theta}_m)_i - (\theta_m^*)_i\right)_{i\in m} \quad (4.8.5)$$

with $\overline{\theta}_m = c\,\widehat{\theta}_m + (1-c)\theta_m^*$ and $0 < c < 1$.
Using $\widehat{\theta}_m \xrightarrow[n\to+\infty]{a.s.} \theta_m^*$ and the ergodic theorem $\frac{1}{n}\,\partial^2_{\theta_i\theta_j} L_n(\theta_m) \xrightarrow[n\to+\infty]{a.s.} F(\theta_m)$ for any $\theta_m \in \Theta_m$ since $\mathbb{E}\big[\big\|\partial^2_{\theta^2}\gamma(\theta, X_0)\big\|_\Theta\big] < \infty$, we obtain:

$$\Big(\frac{1}{n}\,\partial^2_{\theta_i\theta_j} L_n(\overline{\theta}_m)\Big)_{i,j\in m} \xrightarrow[n\to+\infty]{a.s.} F(\theta_m^*). \quad (4.8.6)$$

Finally, by definition of $\widehat{\theta}_m$, $\partial_{\theta_j}\widehat{L}_n(\widehat{\theta}_m) = 0$ for any $j \in m$. As a consequence,

$$\frac{1}{\sqrt{n}}\left(\partial_{\theta_j} L_n(\widehat{\theta}_m)\right)_{j\in m} \xrightarrow[n\to\infty]{\mathcal{P}} 0, \quad (4.8.7)$$

using a Markov Inequality and $\mathbb{E}\big[\frac{1}{\sqrt{n}}\,\|\partial_\theta \widehat{L}_n(\theta) - \partial_\theta L_n(\theta)\|_\Theta\big] \underset{n\to\infty}{\longrightarrow} 0$ established in (5.11) of Bardet and Wintenberger (2009). Considering (4.8.5), (4.8.6) and (4.8.7), and with the central limit theorem satisfied by $\frac{1}{\sqrt{n}}\big(\partial_{\theta_j} L_n(\theta^*_m)\big)_{j\in m}$ provided in Corollary 4.1, this achieves the proof. □

Now, before establishing Proposition 4.1, three technical lemmas can be stated:

Lemma 4.4. *Under the assumptions* **A0**-**A5**, *with* $8/3 < r' \le r/3$ *and* $r' < 2(\delta-1)$ *where* $\delta > 7/2$ *is given in Assumption A5, for any* $m \in \mathcal{M}$, *there exists* $C > 0$ *such as for any* $n \in \mathbb{N}^*$

$$\Big\|\Big(\frac{1}{\sqrt{n}}\,\partial_{\theta_j} L_n(\theta^*_m)\Big)_{j\in m}\Big\|_{r'} \le C. \tag{4.8.8}$$

Proof. First, for any $m \in \mathcal{M}$ and $n \in \mathbb{N}^*$,

$$\begin{aligned}\big\|\big(\partial_{\theta_j} L_n(\theta^*_m)\big)_{j\in m}\big\|^{r'} &\le |m|^{r'/2-1}\sum_{j\in m}\big|\partial_{\theta_j} L_n(\theta^*_m)\big|^{r'} \\ &\le \frac{|m|^{r'/2-1}}{2^{r'}}\sum_{j\in m}\Big|\sum_{t=1}^n \partial_{\theta_j}\gamma(\theta^*_m, X_t)\Big|^{r'}. \end{aligned}\tag{4.8.9}$$

Now, for all $j \in m$, $\big(\partial_{\theta_j}\gamma(\theta^*_m, X_t)\big)_{t\in\mathbb{Z}}$ is a centered (from the proof of Corollary 4.1) stationary $\tau^{(p)}_{\phi^{(j)}_\theta}$-weakly dependent where its coefficients $\big(\tau^{(1)}_{\phi^{(j)}_\theta}(s)\big)_s$ satisfies (4.8.3) (see Lemma 4.3). Moreover, from the proof of Proposition 4.3, $\tau^{(1)}_{\phi^{(j)}_\theta}(s) = O\big(s^{1-\delta}\log s\big)$.

In Proposition 5.5 of Dedecker et al. (2007), since $\mathbb{E}\big[\big|\partial_{\theta_j}\gamma(\theta^*_m, X_0)\big|^{r'}\big] < \infty$ from Lemma 4.6, it has been established that:

$$\mathbb{E}\Big[\Big|\sum_{t=1}^n \partial_{\theta_j}\gamma(\theta^*_m, X_t)\Big|^{r'}\Big] \le C_{r'}\left(M_{r',n} + M_{2,n}^{r'/2}\right)$$

$$\text{where}\quad M_{m,n} := 2n\sum_{i=0}^{n-1}(i+1)^{m-2}\,\tau^{(1)}_{\phi^{(j)}_\theta}(i).$$

Using $8/3 < r' < 2(\delta-1)$ with $\delta > 7/2$, we obtain that

$$M_{r',n} \le C\,n\sum_{i=1}^n i^{r'-1-\delta}\log(i) \le C'\,n^{1+r'-\delta}\log(n) = O\big(n^{r'/2}\big)$$

and $M_{2,n} \le C\,n\sum_{i=1}^n i^{1-\delta}\log(i) \le C''\,n$. As a consequence, there exists $C > 0$ such as for any $n \in \mathbb{N}^*$,

$$\mathbb{E}\Big[\Big|\sum_{t=1}^n \partial_{\theta_j}\gamma(\theta^*_m, X_t)\Big|^{r'}\Big] \le C\,n^{r'/2}. \tag{4.8.10}$$

Then, using (4.8.9) and (4.8.10), the proof is established. □

Lemma 4.5. *Under the assumptions* **A0**-**A5**, *then for any* $m \in \mathcal{M}$, *there exists* $C > 0$ *such as for any* $n \in \mathbb{N}^*$,

$$\Big\|\Big(\frac{1}{\sqrt{n}}\,\partial_{\theta_j} L_n(\widehat{\theta}_m)\Big)_{j\in m}\Big\|_{r/3} \le C.$$

Proof. First, from the definition of $\widehat{\theta}_m$, we have $\partial_{\theta_j}\widehat{L}_n(\widehat{\theta}_m)=0$ for any $j\in m$. Then,

$$\begin{aligned}\Big\|\Big(\frac{1}{\sqrt{n}}\,\partial_{\theta_j}L_n(\widehat{\theta}_m)\Big)_{j\in m}\Big\|_{r/3} &= \Big\|\Big(\frac{1}{\sqrt{n}}\big(\partial_{\theta_j}L_n(\widehat{\theta}_m)-\partial_{\theta_j}\widehat{L}_n(\widehat{\theta}_m)\big)\Big)_{j\in m}\Big\|_{r/3}\\ &\le \frac{|m|^{(r-6)/2r}}{\sqrt{n}}\sum_{j\in m}\Big\|\partial_{\theta_j}L_n(\widehat{\theta}_m)-\partial_{\theta_j}\widehat{L}_n(\widehat{\theta}_m)\Big\|_{r/3}\\ &\le \frac{|m|^{1/2}}{2\sqrt{n}}\sum_{j\in m}\sum_{t=1}^{n}\Big\|\partial_{\theta_j}\gamma(\widehat{\theta}_m,X_t)-\partial_{\theta_j}\widehat{\gamma}(\widehat{\theta}_m,X_t)\Big\|_{r/3}.\end{aligned}$$

From the proof of Lemma 2 in Bardet et al. (2020b), there exists $C>0$ such as

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\theta\in\Theta}\Big\|\partial_{\theta_j}\gamma(\theta,X_t)-\partial_{\theta_j}\widehat{\gamma}(\theta,X_t)\Big\|^{r/3}\Big]\\ &\qquad\le C\Big(\sum_{k\ge t}\alpha_k(f_\theta,\Theta)+\alpha_k(M_\theta,\Theta)+\alpha_k(\partial f_\theta,\Theta)+\alpha_k(\partial M_\theta,\Theta)\Big)^{r/3}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Big\|\Big(\frac{1}{\sqrt{n}}\,\partial_{\theta_j}L_n(\widehat{\theta}_m)\Big)_{j\in m}\Big\|_{r/3} &\le C\,\frac{|m|^{3/2}}{2\sqrt{n}}\sum_{t=1}^{n}\sum_{k\ge t}\big(\alpha_k(f_\theta,\Theta)+\alpha_k(M_\theta,\Theta)+\alpha_k(\partial f_\theta,\Theta)+\alpha_k(\partial M_\theta,\Theta)\big)\\ &\le \frac{C'}{\sqrt{n}}\sum_{t=1}^{n}\sum_{j\ge t}j^{-\delta}\le\frac{C''}{\sqrt{n}}\sum_{t=1}^{n}t^{1-\delta}\le C''',\end{aligned}$$

with $C'>0$, $C''>0$ and $C'''>0$ and where the last inequality holds since $\delta>7/2$ under Assumption A5. □

Lemma 4.6. *Under the assumptions* **A0**-**A5**, *for any* $m\in\mathcal{M}$ *and any* $\theta\in\Theta_m$,

$$\Big\|\big(\partial_{\theta_j}\gamma(\theta,X_0)\big)_{j\in m}\Big\|_{r/3}<\infty. \tag{4.8.11}$$

Proof. For $j\in m$, we have for any $\theta\in\Theta_m$,

$$\partial_{\theta_j}\gamma(\theta,X_0)=-2\,(M_\theta^0)^{-2}(X_0-f_\theta^0)\,\partial_{\theta_j}f_\theta^0-2\,(M_\theta^0)^{-3}(X_0-f_\theta^0)^2\partial_{\theta_j}M_\theta^0+2\,(M_\theta^0)^{-1}\partial_{\theta_j}M_\theta^0.$$

Therefore with Assumption A3 and Minkowski Inequality,

$$\begin{aligned}\big\|\big(\partial_{\theta_j}\gamma(\theta,X_0)\big)_{j\in m}\big\|_{r/3}\le\frac{2}{\underline{h}^{3/2}}\Big(&\underline{h}^{1/2}\,\big\|\big(\partial_{\theta_j}f_\theta^0\big)\big(X_0-f_\theta^0\big)\big\|_{r/3}\\ &+\big\|\big(\partial_{\theta_j}M_\theta^0\big)\,\big|X_0-f_\theta^0\big|^2\big\|_{r/3}+\underline{h}\,\big\|\big(\partial_{\theta_p}M_\theta^0\big)\big\|_{r/3}\Big).\end{aligned}$$

Now, applying the Hölder Inequality, we obtain that there exists $C>0$ such that for any $\theta\in\Theta_m$,

$$\begin{aligned}\mathbb{E}\Big[\big|\partial_{\theta_p}\gamma(\theta^*,X_0)\big|^{r/3}\Big]\le C\Big(&\big\|\partial_{\theta_j}f_\theta^0\big\|_{2r/3}\,\big\|X_0-f_\theta^0\big\|_{2r/3}\\ &+\big\|\partial_{\theta_j}M_\theta^0\big\|_r\,\big\|X_0-f_\theta^0\big\|_r^2+\big\|\partial_{\theta_p}M_\theta^0\big\|_{r/3}\Big).\end{aligned}\tag{4.8.12}$$

Using Assumption A0 and A5 and the proof of Lemma 1 in Bardet and Wintenberger (2009), all the right side terms in (4.8.12) are finite for any $\theta\in\Theta_m$ and this achieves the proof. □

Then Proposition 4.1 can be established:

Proof of Proposition 4.1. From (4.8.5) and (4.8.6) with $F(\theta_m^*)$ is a positive definite matrix, we know that for n large enough,

$$\begin{aligned}
&\big\|\sqrt{n}\,\big((\widehat{\theta}_m)_i-(\theta_m^*)_i\big)_{i\in m}\big\|_{r'}\\
&\qquad=\Big\|\Big(\big(\frac{1}{n}\,\partial^2_{\theta_i\theta_j}L_n(\overline{\theta}_m)\big)_{i,j\in m}\Big)^{-1}\times\frac{1}{\sqrt{n}}\,\big(\partial_{\theta_j}L_n(\widehat{\theta}_m)-\partial_{\theta_j}L_n(\theta_m^*)\big)_{j\in m}\Big\|_{r'}. \qquad (4.8.13)
\end{aligned}$$

Therefore, using Hölder and Minkowski inequalities, we obtain:

$$\begin{aligned}
\big\|\sqrt{n}\,\big((\widehat{\theta}_m)_i-(\theta_m^*)_i\big)_{i\in m}\big\|_{r'} &\le \Big\|\Big(\big(\frac{1}{n}\,\partial^2_{\theta_i\theta_j}L_n(\overline{\theta}_m)\big)_{i,j\in m}\Big)^{-1}\Big\|_{\frac{rr'}{r-3r'}}\\
&\qquad\times\Big\|\frac{1}{\sqrt{n}}\,\big(\partial_{\theta_j}L_n(\widehat{\theta}_m)-\partial_{\theta_j}L_n(\theta_m^*)\big)_{j\in m}\Big\|_{\frac{r}{3}}\\
&\le \Big\|\Big(\big(\frac{1}{n}\,\partial^2_{\theta_i\theta_j}L_n(\overline{\theta}_m)\big)_{i,j\in m}\Big)^{-1}\Big\|_{\frac{rr'}{r-3r'}}\\
&\qquad\times\Big(\Big\|\frac{1}{\sqrt{n}}\,\big(\partial_{\theta_j}L_n(\widehat{\theta}_m)\big)_{j\in m}\Big\|_{\frac{r}{3}}+\Big\|\frac{1}{\sqrt{n}}\,\big(\partial_{\theta_j}L_n(\theta_m^*)\big)_{j\in m}\Big\|_{\frac{r}{3}}\Big).
\end{aligned}$$

Now using Assumption A4, Lemmas 4.4 and 4.5, we deduce (4.3.7). □

4.8.2 Proofs of Section 4.4

Proof of Lemma 4.1. **1.** From the assumptions, the function $R:\ \theta\in\Theta\mapsto R(\theta)$ is a $\mathcal{C}^2(\Theta)$ function and the Hessian matrix $\partial_{\theta^2}R=F$ is a definite positive matrix. Therefore, from a Taylor-Lagrange expansion:

$$\begin{aligned}
n\big(R(\widehat{\theta}_m)-R(\theta_m^*)\big) &= n\Big(R(\theta_m^*)+\big(\widehat{\theta}_m-\theta_m^*\big)^\top\partial_\theta R(\theta_m^*)\\
&\qquad+\frac{1}{2}\,\big(\widehat{\theta}_m-\theta_m^*\big)^\top\partial_{\theta^2}R(\overline{\theta})\,\big(\widehat{\theta}_m-\theta_m^*\big)-R(\theta_m^*)\Big)\\
&= \frac{1}{2}\,\big(\sqrt{n}(\widehat{\theta}_m-\theta_m^*)\big)^\top\partial_{\theta^2}R(\overline{\theta})\,\big(\sqrt{n}(\widehat{\theta}_m-\theta_m^*)\big), \qquad (4.8.14)
\end{aligned}$$

with $\overline{\theta}=\theta_m^*+c\,\big(\widehat{\theta}_m-\theta_m^*\big)\in\Theta_m$ since $c\in[0,1]$. Using Lemma 4 of Bardet and Wintenberger (2009) and continuous mapping Theorem, we deduce that:

$$\partial_{\theta^2}R(\overline{\theta})=F(\overline{\theta})\xrightarrow[n\to\infty]{\mathcal{P}}-2\,F(\theta_m^*)\quad\text{and}\quad G(\overline{\theta})\xrightarrow[n\to\infty]{\mathcal{P}}G(\theta_m^*). \qquad (4.8.15)$$

Moreover, using the asymptotic normality of $\widehat{\theta}_m$ established in Bardet and Wintenberger (2009) and Bardet et al. (2020b), we have:

$$\sqrt{n}((\widehat{\theta}_m)_i-(\theta_m^*)_i)_{i\in m}\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}\Big(0\,,\ \big(F(\theta_m^*)\big)^{-1}G(\theta_m^*)\big(F(\theta_m^*)\big)^{-1}\Big). \qquad (4.8.16)$$

As a consequence, with $Z_n=\big(G(\overline{\theta})\big)^{-1/2}F(\overline{\theta})\sqrt{n}((\widehat{\theta}_m)_i-(\theta_m^*)_i)_{i\in m}\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}(0\,,\,I_{|m|})$, we have

$$\begin{aligned}
n\big(R(\widehat{\theta}_m)-R(\theta_m^*)\big) &= -Z_n^\top\big(G(\overline{\theta})\big)^{1/2}\big(F(\overline{\theta})\big)^{-1}F(\overline{\theta})\big(F(\overline{\theta})\big)^{-1}\big(G(\overline{\theta})\big)^{1/2}Z_n\\
&= -Z_n^\top\big(G(\overline{\theta})\big)^{1/2}\big(F(\overline{\theta})\big)^{-1}\big(G(\overline{\theta})\big)^{1/2}Z_n.
\end{aligned}$$

Define $U^*(m) := -Z^\top \big(G(\theta_m^*)\big)^{1/2}\big(F(\theta_m^*)\big)^{-1}\big(G(\theta_m^*)\big)^{1/2} Z$ where $Z \overset{\mathcal{D}}{\sim} \mathcal{N}(0\,,\, I_{|m|})$. Then using (4.8.15) we obtain

$$n\big(R(\widehat{\theta}_m) - R(\theta_m^*)\big) \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} U^*(m).$$

The computation of the expectation of U_m^* follows from

$$\begin{aligned}\mathbb{E}\big[U_m^*\big] &= \mathbb{E}\big[\mathrm{Trace}\big(U_m^*\big)\big] = -\,\mathrm{Trace}\Big(\big(G(\theta_m^*)\big)^{1/2}\big(F(\theta_m^*)\big)^{-1}\big(G(\theta_m^*)\big)^{1/2}\Big)\\ &= -\,\mathrm{Trace}\Big(\big(F(\theta_m^*)\big)^{-1} G(\theta_m^*)\Big).\end{aligned}$$

Finally, for establishing $\mathbb{E}\big[n\big(R(\widehat{\theta}_m) - R(\theta_m^*)\big)\big] \underset{n\to\infty}{\longrightarrow} \mathbb{E}\big[U_m^*\big]$, we have to prove that there exists $n_0 \in \mathbb{N}$ such as

$$\sup_{n\geq n_0} \mathbb{E}\big[n\big|R(\widehat{\theta}_m) - R(\theta_m^*)\big|\big] < \infty. \tag{4.8.17}$$

Indeed, from (4.8.14), we have:

$$\begin{aligned} n\,\big|R(\widehat{\theta}_m) - R(\theta_m^*)\big| &\leq \frac{1}{2}\sup_{\theta\in\Theta}\big\|\partial_{\theta^2}^2 R(\theta)\big\|\,\big\|\sqrt{n}(\widehat{\theta}_m - \theta_m^*)\big\|^2\\ &\Longrightarrow \quad \mathbb{E}\big[n\big|R(\widehat{\theta}_m) - R(\theta_m^*)\big|\big] \leq \frac{\lambda_{\max}}{2}\,\mathbb{E}\big[\big\|\sqrt{n}(\widehat{\theta}_m - \theta_m^*)\big\|^2\big], \end{aligned}\tag{4.8.18}$$

since there exists $\lambda_{\max} < \infty$ such as $\big\|\partial_{\theta^2} R(\theta)\big\| \leq \lambda_{\max}$ for any $\theta \in \Theta$ from Assumption $A_2(f_\theta, \Theta)$ and $A_2(M_\theta, \Theta)$ where Θ is a compact set.
Using Proposition 4.1, we know that

$$\sup_{n\in\mathbb{N}^*}\big\|\sqrt{n}\,\big(\widehat{\theta}_m - \theta_m^*\big)\big\|_2 < \infty.$$

Finally using (4.8.18), we deduce (4.8.17).

2. As in the proof of **1.**, we use a Taylor-Lagrange expansion of $\widehat{\gamma}_n(\theta_m^*)$ around $\widehat{\theta}_m$ since $\partial_\theta \widehat{\gamma}_n(\widehat{\theta}_m) = 0$. Then,

$$n\,\big(\widehat{\gamma}_n(\theta_m^*) - \widehat{\gamma}_n(\widehat{\theta}_m)\big) = \frac{1}{2}\,\sqrt{n}(\widehat{\theta}_m - \theta_m^*)^\top\,\big(\partial_{\theta^2}^2\widehat{\gamma}_n(\overline{\theta}_m)\big)\,\sqrt{n}\big(\widehat{\theta}_m - \theta_m^*\big).$$

But using $\widehat{\theta}_m \underset{n\to+\infty}{\xrightarrow{a.s.}} \theta_m^*$ and $\mathbb{E}\Big[\Big\|\frac{1}{n}\,L_n(\theta) - \frac{1}{n}\,\widehat{L}_n(\theta)\Big\|_\Theta\Big] \underset{n\to\infty}{\longrightarrow} 0$, we have

$$\big(\partial_{\theta^2}^2\widehat{\gamma}_n(\overline{\theta}_m)\big) \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} -\,2\,F(\theta_m^*).$$

Therefore, using the same reasoning as in 1., we deduce that

$$n\,\big(\widehat{\gamma}_n(\theta_m^*) - \widehat{\gamma}_n(\widehat{\theta}_m)\big) \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} U^*(m).$$

With Hölder Inequality and using $8/3 < r'$ defined in Proposition 4.1, we obtain for n large enough

$$\begin{aligned}\mathbb{E}\big[n\,\big(\widehat{\gamma}_n(\theta_m^*) - \widehat{\gamma}_n(\widehat{\theta}_m)\big)\big] &\leq \big\|\big(\partial_{\theta^2}^2\widehat{\gamma}_n(\overline{\theta}_m)\big)^{1/2}\,\sqrt{n}(\widehat{\theta}_m - \theta_m^*)\big\|_2^2\\ &\leq \Big\|\big(\partial_{\theta^2}^2\widehat{\gamma}_n(\overline{\theta}_m)\big)\Big\|_{\frac{r'}{r'-2}}\,\big\|\sqrt{n}(\widehat{\theta}_m - \theta_m^*)\big\|_{r'}^2. \end{aligned}\tag{4.8.19}$$

Finally, with Proposition 4.1 and Lemma 4 of Bardet and Wintenberger (2009), we have $\sup_{n\in\mathbb{N}^*}\mathbb{E}\big[\big\|\partial^2_{\theta^2}\widehat{\gamma}_n(\overline{\theta}_m)\big\|^4_\Theta\big]<\infty$ and $\frac{r'}{r'-2}\leq 4$ since $r'>8/3$, and therefore

$$\sup_{n\in\mathbb{N}^*}\mathbb{E}\big[n\left(\widehat{\gamma}_n(\theta^*_m)-\widehat{\gamma}_n(\widehat{\theta}_m)\right)\big]<\infty,$$

which concludes the proof. □

Proof of Proposition 1.5. The proof of this proposition can be deduced from

$$\mathbb{E}\big[n\,I_3(m)\big]=\mathbb{E}\big[n\left(R(\theta^*_m)-\widehat{\gamma}_n(\theta^*_m)\right)\big]=v^*_n \tag{4.8.20}$$

for any $m\in\mathcal{M}$. For establishing (4.8.20), we begin by

$$I_3(m)=\big(R(\theta^*_m)-\gamma_n(\theta^*_m)\big)+\big(\gamma_n(\theta^*_m)-\widehat{\gamma}_n(\theta^*_m)\big):=I_{31}(m)+I_{32}(m). \tag{4.8.21}$$

Firstly, since $\mathbb{E}\big[\gamma(\theta^*_m,X_0)\big]=R(\theta^*_m)$ and $(X_t)_{t\in\mathbb{Z}}$ is a stationary times series, then for any $n\in\mathbb{N}^*$,

$$\mathbb{E}\big[\gamma_n(\theta^*_m)\big]=\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\big[\gamma(\theta^*_m,X_t)\big]=R(\theta^*_m)\quad\Longrightarrow\quad\mathbb{E}\big[I_{31}(m)\big]=0. \tag{4.8.22}$$

Secondly, from Assumption A0 and Bardet and Wintenberger (2009), there exists $C>0$ such that for any $t\geq 1$

$$\mathbb{E}\big[\big\|\gamma(\theta,X_t)-\widehat{\gamma}(\theta,X_t)\big\|_\Theta\big]\leq C\sum_{s\geq t}\big(\alpha_k(f_\theta,\Theta)+\alpha_k(M_\theta,\Theta)\big).$$

Therefore,

$$\begin{aligned}\mathbb{E}\big[\big\|\gamma_n(\theta^*_m)-\widehat{\gamma}_n(\theta^*_m)\big\|_\Theta\big]&\leq\frac{C}{n}\sum_{t=1}^{n}\sum_{s\geq t}\big(\alpha_k(f_\theta,\Theta)+\alpha_k(M_\theta,\Theta)\big)\\&\leq\frac{C}{n}\sum_{t=1}^{n}t^{1-\delta}\leq\frac{C'}{n},\end{aligned} \tag{4.8.23}$$

since $\delta>7/2$ from Assumption A5. Moreover, for any $m\in\mathcal{M}$ such as $m^*\subset m$, we have $\gamma_n(\theta^*_m)-\widehat{\gamma}_n(\theta^*_m)=\gamma_n(\theta^*_m)-\widehat{\gamma}_n(\theta^*_m)$. Using this and (4.8.23) we deduce that for any $m\in\mathcal{M}$, there exists a bounded sequence $(v^*_n)_{n\in\mathbb{N}^*}$ not depending on m satisfying

$$\mathbb{E}\big[I_{32}(m)\big]=\frac{v^*_n}{n}. \tag{4.8.24}$$

Using also Lemma 4.1, this implies the asymptotic behavior of $\mathbb{E}\big[\mathrm{pen}_{id}(m)\big]$. □

Now we establish a preliminary lemma that is an important step towards the proof of Theorem 4.3.

Lemma 4.7. *Let pen* : $m\in\mathcal{M}_n\mapsto\mathrm{pen}(m)\in\mathbb{R}_+$. *Then*

$$\ell(\widehat{\theta}_{\widehat{m}_{\mathrm{pen}}},\theta^*)\leq\min_{m\in\mathcal{M}}\big\{\ell(\widehat{\theta}_m,\theta^*)\big\}+\big(\mathrm{pen}(\widehat{m}_{id})-\mathrm{pen}(\widehat{m}_{\mathrm{pen}})\big)-\big(\mathrm{pen}_{id}(\widehat{m}_{id})-\mathrm{pen}_{id}(\widehat{m}_{\mathrm{pen}})\big). \tag{4.8.25}$$

Proof. By definition, for any $m \in \mathcal{M}$,

$$\widehat{C}_{\mathrm{pen}_{id}}(m) = R(\widehat{\theta}_m) = \ell(\widehat{\theta}_m, \theta^*) + R(\theta^*). \tag{4.8.26}$$

As a consequence,

$$\min_{m\in\mathcal{M}} \big\{\ell(\widehat{\theta}_m, \theta^*)\big\} = \ell(\widehat{\theta}_{\widehat{m}_{id}}, \theta^*) = \min_{m\in\mathcal{M}} \big\{\widehat{C}_{\mathrm{pen}_{id}}(m)\big\} + R(\theta^*). \tag{4.8.27}$$

For any $m \in \mathcal{M}$, we also have

$$\widehat{C}_{\mathrm{pen}}(m) = \widehat{C}_{\mathrm{pen}_{id}}(m) + \mathrm{pen}(m) - \mathrm{pen}_{id}(m).$$

By definition of $\widehat{m}_{\mathrm{pen}}$, we have $\widehat{C}_{\mathrm{pen}}(\widehat{m}_{\mathrm{pen}}) \le \widehat{C}_{\mathrm{pen}}(\widehat{m}_{\mathrm{pen}_{id}})$. Therefore,

$$\begin{aligned} \widehat{C}_{\mathrm{pen}}(\widehat{m}_{\mathrm{pen}}) &\le \widehat{C}_{\mathrm{pen}_{id}}(\widehat{m}_{\mathrm{pen}_{id}}) + \mathrm{pen}(\widehat{m}_{\mathrm{pen}_{id}}) - \mathrm{pen}_{id}(\widehat{m}_{\mathrm{pen}_{id}}) \\ C_{\mathrm{pen}_{id}}(\widehat{m}_{\mathrm{pen}}) + \mathrm{pen}(\widehat{m}_{\mathrm{pen}}) - \mathrm{pen}_{id}(\widehat{m}_{\mathrm{pen}}) &\le \widehat{C}_{\mathrm{pen}_{id}}(\widehat{m}_{\mathrm{pen}_{id}}) + \mathrm{pen}(\widehat{m}_{\mathrm{pen}_{id}}) - \mathrm{pen}_{id}(\widehat{m}_{\mathrm{pen}_{id}}). \end{aligned}$$

By replacing $\widehat{C}_{\mathrm{pen}_{id}}(m)$ by $\ell(\widehat{\theta}_m, \theta^*)$ following (4.8.26) and using (4.8.27), then (4.8.25) is established. □

Proof of Theorem 4.2. Let $\mathcal{M}^* = \big\{m \in \mathcal{M},\ m^* \subset m\big\}$ and $\mathcal{M}' = \mathcal{M} \setminus \mathcal{M}^*$. Let $m \in \mathcal{M}'$. We have:

$$\begin{aligned} \mathbb{P}\big(\widehat{m}_{\mathrm{pen}} = m\big) &\le \mathbb{P}\big(\widehat{C}_{\mathrm{pen}}(m) \le \widehat{C}_{\mathrm{pen}}(m^*)\big) \\ &\le \mathbb{P}\Big\{\big(\widehat{\gamma}_n(\widehat{\theta}_m) - \widehat{\gamma}_n(\widehat{\theta}_{m^*})\big) \le \mathrm{pen}(m^*) - \mathrm{pen}(m)\big)\Big\} \\ &\le \mathbb{P}\Big\{n\big(\widehat{\gamma}_n(\widehat{\theta}_m) - \widehat{\gamma}_n(\theta^*_m)\big) + n\big(\widehat{\gamma}_n(\theta^*_m) - R(\theta^*_m)\big) + n\big(R(\theta^*) - \widehat{\gamma}_n(\theta^*)\big) \\ &\qquad + n\big(\widehat{\gamma}_n(\theta^*) - \widehat{\gamma}_n(\widehat{\theta}_{m^*})\big) \le n\big(R(\theta^*) - R(\theta^*_m)\big) + n\big(\mathrm{pen}(m^*) - \mathrm{pen}(m)\big)\Big\} \\ &\le \mathbb{P}\Big\{Z_1 + Z_2 + Z_3 + Z_4 + Z5 \le -2\,n\,DK_L(\theta^*\|\theta^*_m)\Big\} \end{aligned}$$

with $Z_5 = n\big(\mathrm{pen}(m) - \mathrm{pen}(m^*)\big)$ and with $R(\theta^*) - R(\theta^*_m) = -2\,DK_L(\theta^*\|\theta^*_m) < 0$ since $m \not\subset m^*$ from Bardet et al. (2020b). Now, using $\mathbb{P}(Z_1 + \cdots + Z_5 \le c) \le \mathbb{P}(Z_1 \le c/5) + \cdots + \mathbb{P}(Z_5 \le c/5)$ for any random variables Z_i and real number c, we obtain:

$$\mathbb{P}\big(\widehat{m}_{\mathrm{pen}} = m\big) \le \sum_{i=1}^{5} \mathbb{P}\big(Z_i \le c_n\big), \tag{4.8.28}$$

where $c_n = -\frac{2}{5}\,n\,DK_L(\theta^*\|\theta^*_m)$.
Let $Z_1 := n\big(\widehat{\gamma}_n(\widehat{\theta}_m) - \widehat{\gamma}_n(\theta^*_m)\big)$. Following the same computations than in (4.8.19), with $8/3 < r' \le r/3$ and $r' < 2(\delta - 1)$ defined in Proposition 4.1, and Hölder Inequality,

$$\begin{aligned} \mathbb{E}\big[\big|Z_1\big|^{\frac{3r'}{8}}\big] &\le \big\|\big(\partial^2_{\theta^2}\widehat{\gamma}_n(\overline{\theta}_m)\big)^{1/2}\sqrt{n}(\widehat{\theta}_m - \theta^*_m)\big\|^{\frac{3r'}{4}}_{\frac{3r'}{4}} \\ &\le \Big\|\big(\partial^2_{\theta^2}\widehat{\gamma}_n(\overline{\theta}_m)\big)\Big\|_{\frac{3r'}{2}} \big\|\sqrt{n}(\widehat{\theta}_m - \theta^*_m)\big\|^{\frac{3r'}{4}}_{r'}. \end{aligned}$$

Therefore, using Proposition 4.1,

$$\mathbb{P}\big(Z_1 \le c_n\big) \le \mathbb{P}\Big(\big|Z_1\big|^{\frac{3r'}{8}} \ge \big(|c_n|\big)^{\frac{3r'}{8}}\Big) \le \mathbb{E}\big[\big|Z_1\big|^{\frac{3r'}{8}}\big]\frac{1}{|c_n|^{\frac{3r'}{8}}}$$

$$\Longrightarrow \quad \mathbb{P}\big(Z_1 \le c_n\big) = O\Big(\frac{1}{n^{\frac{3r'}{8}}}\Big) = o\big(\frac{1}{n}\big), \tag{4.8.29}$$

since $3r'/8 > 1$. The same kind of computations can also be done for $Z_4 := n\big(\widehat{\gamma}_n(\theta^*) - \widehat{\gamma}_n(\widehat{\theta}_{m^*})\big)$ and we also obtain $\mathbb{P}\big(Z_4 \leq c_n\big) = o\big(\frac{1}{n}\big)$.

Consider now $Z_2 := n\big(\widehat{\gamma}_n(\theta^*_m) - R(\theta^*_m)\big)$. Then,

$$\mathbb{E}\big[\big|Z_2\big|^{8/3}\big] \leq 2^{5/3}\Big(\mathbb{E}\big[\big\|\widehat{L}_n(\theta) - L_n(\theta)\big\|_\Theta^{8/3}\big] + n^{8/3}\,\mathbb{E}\big[\big|\sum_{k=1}^n \big(\gamma(\theta^*_m, X_k) - R(\theta^*_m)\big)\big|^{8/3}\big]\Big).$$

Using Bardet and Wintenberger (2009), we know that $\sup_{n\in\mathbb{N}^*} E\big[\big\|\widehat{L}_n(\theta) - L_n(\theta)\big\|_\Theta^{8/3}\big] < \infty$ from Assumption A5 and since $\delta > 7/2 > 2$. Now, consider $Y_k := \gamma(\theta^*_m, X_k) - R(\theta^*_m)$. Then, $(Y_k)_{k\in\mathbb{Z}}$ is a stationary time series, τ_Y-weakly dependent because, using the same type of arguments as in the proof of Lemma 4.3, we have:

$$\tau_Y(s) \leq \sum_{\ell=1}^{\infty}\big(\alpha_\ell(f_\theta, \Theta) + \alpha_\ell(M_\theta, \Theta)\big)\,\lambda_{s+1-\ell},$$

with λ defined in Lemma 4.2. Therefore, using Assumption A5, we also have $\tau_Y(s) = O\big(s^{\delta-1}\log(s)\big)$, with $\delta > 7/2$. Now, using the same type of arguments as in the proof of Lemma 4.4,

$$\mathbb{E}\big[\big|\sum_{k=1}^n Y_k\big|^{8/3}\big] \leq C_{8/3}\,\big(M_{8/3,n} + M_{2,n}^{4/3}\big),$$

and $M_{2,n} \leq C\,n$ while $M_{8/3,n} \leq C\,n\sum_{i=1}^n i^{8/3-1-\delta}\log(i) = o\big(n^{4/3}\big)$. Therefore, there exists $C > 0$ such that for any $n \in \mathbb{N}^*$,

$$\mathbb{E}\big[\big|\sum_{k=1}^n Y_k\big|^{8/3}\big] \leq C\,n^{4/3}.$$

Finally, we deduce that there exists $C > 0$ such that for any $n \in \mathbb{N}^*$,

$$\mathbb{E}\big[\big|Z_2\big|^{8/3}\big] \leq C\,n^{4/3}. \tag{4.8.30}$$

This result and Markov Inequality imply,

$$\begin{aligned}\mathbb{P}\big(Z_2 \leq c_n\big) \leq \mathbb{P}\Big(|Z_2|^{8/3} \geq \big(|c_n|\big)^{8/3}\Big) \leq \mathbb{E}\big[\big|Z_2\big|^{8/3}\big]\,\frac{1}{|c_n|^{8/3}}\\ \Longrightarrow \quad \mathbb{P}\big(Z_2 \leq c_n\big) = O\Big(n^{4/3}\,\frac{1}{|c_n|^{8/3}}\Big) = O\big(\frac{1}{n^{4/3}}\big),\end{aligned} \tag{4.8.31}$$

We obtain the same bound for $Z_3 := n\big(R(\theta^*) - \widehat{\gamma}_n(\theta^*)\big)$.
Finally using the assumption (1.3.20), we have:

$$\begin{aligned}n\,\mathbb{P}\big(Z_5 \leq c_n\big) = \mathbb{P}\Big(\big(\mathrm{pen}(m) - \mathrm{pen}(m^*)\big) \leq -\frac{2}{5}\,DK_L(\theta^*\|\theta^*_m)\Big)\\ \leq \mathbb{P}\Big(\mathrm{pen}(m^*) \geq \frac{2}{5}\,DK_L(\theta^*\|\theta^*_m)\Big) \underset{n\to\infty}{\longrightarrow} 0.\end{aligned} \tag{4.8.32}$$

By this way, (4.4.6) is established. □

Proof of Theorem 4.3. The proof is mainly based on Lemma 4.7. From the proof of Lemma 4.1, we deduce that for any $m \in \mathcal{M}$, there exists two positive random variable $Y(m)$ and

$Z(m)$ such as $n\left|I_1(m)+I_2(m)\right| \leq Y(m)$ and $n\left|I_3(m)\right| \leq Z$ for any $n \in \mathbb{N}^*$. Moreover, $Y(m)$ and Z have bounded expectations. Therefore, using Markov Inequality and since $\mathcal{M}$ is supposed to be a finite family of models, for any $\varepsilon > 0$ there exists $K'_\varepsilon > 0$ such as

$$\limsup_{n\to\infty} \max_{m\in\mathcal{M}} \mathbb{P}\Big(n\,\mathrm{pen}_{id}(m) \geq K'_\varepsilon\Big) \leq \varepsilon.$$

Therefore, using this inequality and (4.4.7), we deduce that for any $\varepsilon > 0$ there exist $M_\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}^*$ such that for any $n \geq N_\varepsilon$,

$$\mathbb{P}\Big(n\left|\big(\mathrm{pen}(\widehat{m}_{id}) - \mathrm{pen}(\widehat{m}_{\mathrm{pen}})\big) - \big(\mathrm{pen}_{id}(\widehat{m}_{id}) - \mathrm{pen}_{id}(\widehat{m}_{\mathrm{pen}})\big)\right| \leq M_\varepsilon\Big) \geq 1-\varepsilon. \quad (4.8.33)$$

The proof of (4.4.8) is now completed from (4.8.25) of Lemma 4.7 and (4.8.33). □

Proof of Theorem 4.4. We first verify conditions (C1) and (C2) of Chen (1985) that are sufficient to imply Conditions (i), (ii) and (iii) of Kass et al. (1990). Condition (C1) requires that $\widehat{\sigma}_n$ the largest eigenvalue of $\big(-\big(\partial^2_{\theta_i\theta_j}\widehat{L}_n(\widehat{\theta}_m)\big)_{i,j\in m}\big)^{-1}$ satisfies $\widehat{\sigma}_n \xrightarrow[n\to+\infty]{a.s.} 0$, which is satisfied since it was already established that $\frac{1}{n}\big(\partial^2_{\theta_i\theta_j}\widehat{L}_n(\widehat{\theta}_m)\big)_{i,j\in m} \xrightarrow[n\to+\infty]{a.s.} F(\theta^*_m)$ and $F(\theta^*_m)$ is assumed to be a negative definite matrix. Moreover, condition (C2) is also satisfied because $\theta_m \in \Theta_m \mapsto \big(\partial^2_{\theta_i\theta_j}\widehat{L}_n(\theta_m)\big)_{i,j\in m}$ and $\theta_m \in \Theta_m \mapsto \big(\big(\partial^2_{\theta_i\theta_j}\widehat{L}_n(\theta_m)\big)_{i,j\in m}\big)^{-1}$ are continuous functions for n large enough. Therefore, using $h_n = -\frac{1}{n}\widehat{L}_n$, the assumptions of Theorem 1 of Kass et al. (1990) are satisfied and this implies that:

$$\begin{aligned}\int_{\Theta_m} b_m(\theta)\,\exp\big(\widehat{L}_n(\theta)\big)\,d\theta &= \exp\big(\widehat{L}_n(\widehat{\theta}_m)\big)\,(2\,\pi)^{|m|/2} \\ &\quad\times \det\Big(n\big(-\frac{1}{n}\partial^2_{\theta_i\theta_j}\widehat{L}_n(\widehat{\theta}_m)\big)_{i,j\in m}\Big)^{-1/2}\Big(b_m(\widehat{\theta}_m)+O(n^{-1})\Big) \quad a.s.\end{aligned}$$

As a consequence, we have:

$$\begin{aligned}\widehat{S}_n(m,X) &= -\log(|\mathcal{M}|) + \log\Big[\int_{\Theta_m} b_m(\theta)\,\exp\big(\widehat{L}_n(\theta)\big)\,d\theta\Big] \\ &= \widehat{L}_n(\widehat{\theta}_m) - \frac{\log(n)}{2}\,|m| + \log\big(b_m(\widehat{\theta}_m)\big) \\ &\qquad + \frac{\log(2\pi)}{2}\,|m| - \frac{1}{2}\,\log\Big[\det\big(-\widehat{F}_n(m)\big)\Big] - \log(|\mathcal{M}|) + O(n^{-1}) \quad a.s.\end{aligned}$$

□

5

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

The content of this chapter is contained in the submitted preprint, Kare Kamila. "*Data Driven Model Selection for Same-Realization Predictions in Autoregressive Processes*" https://hal.archives-ouvertes.fr/hal-03169343/document.

Abstract

This paper is about the one-step ahead prediction of the future of observations drawn from an infinite-order autoregressive AR(∞) process. It aims to design penalties (fully data driven) ensuring that the selected model verifies the efficiency property but in the non asymptotic framework. We show that the excess risk of the selected estimator enjoys the best bias-variance trade-off over the considered collection. To achieve these results, we needed to overcome the dependence difficulties by following a classical approach which consists in restricting to a set where the empirical covariance matrix is equivalent to the

theoretical one. We show that this event happens with probability larger than $1 - c_0/n^2$ with $c_0 > 0$. The proposed data driven criteria are based on the minimization of the penalized criterion akin to the Mallows's C_p.

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

5.1 Introduction

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

$$X_t = f^*\big((X_{t-i})_{i\in\mathbb{N}^*}\big) + \sigma\,\xi_t \quad \text{for any } t \in \mathbb{Z}. \tag{5.1.1}$$

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

The problem is to estimate the function f^* using these observations. The process (5.1.1) is a particular case of the general class of affine causal process studied in Doukhan and Wintenberger (2008) and Bardet and Wintenberger (2009). The study of this type of process more often requires the classical regularity condition on the function f^*, which are not restrictive at all and remain valid in various time series models. This condition can be stated as follows:

$$\sum_{k=1}^{\infty}\left(\sup_{x\in\mathbb{R}^\infty}\Big|\frac{\partial}{\partial x_k}f^*(x)\Big|\right) < 1, \tag{5.1.2}$$

provided that that f^* admits partial derivatives on $\mathbb{R}^{\mathbb{N}}$. Under (5.1.2) and if the noise ξ_0 admits r-order moments, Doukhan and Wintenberger (2008) showed that there exists a stationary, mixing and ergodic solution to (5.1.1) admitting r-order moments.
Moreover, Bardet and Wintenberger (2009) studied the consistency and the asymptotic normality of the QMLE of $\theta^* = (\theta_i^*)_{i\in\mathbb{N}}$ in the case $f^* = f_{\theta^*}$.

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

$$f^*(X_{t-1}, X_{t-2}, \ldots) = f_{\theta^*}(X_{t-1}, X_{t-2}, \ldots) = \sum_{i=1}^{\infty}\theta_i^* X_{t-i}. \tag{5.1.3}$$

For such processes, condition (5.1.2) becomes

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

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

Henceforth, let observations $(X_1, X_2, \ldots, X_n)$ be a trajectory of the solution $X := (X_t)_{t\in\mathbb{Z}}$ of (5.1.1) verifying **A1**. The goal of this paper is to predict the next value X_{n+1}. In fact, if θ^* were known, a simple prediction of X_{n+1} could be $f_{\theta^*}(X_n, X_{n-1}, \ldots)$ setting $X_t = 0$ for all $t < 0$. However, θ^* is generally unknown and it is impossible to provide a direct estimator since its coordinate are infinite. It is classical to identify a 'good' finite-dimensional model based on the data which can be done by sieve estimation where only a finite number of $\{\theta_i^*\}_{i=1}^K$ is estimated and letting K grows as the sample size increases. A usual approach to this is model selection and the goal is to provide a model with the prediction error as small as the oracle's one.

This question has already been addressed in the literature. Shibata (1980) was the first to tackle this issue. He proved that Akaike criterion is *asymptotically efficient* in the sense that the selected model achieves a smaller one-step mean squared error of prediction when it is fitted to predict an independent realization of the same process. Following Shibata's asymptotically setting, Ing and Wei (2003) and Ing et al. (2005) extended this result for same realization predictions. Indeed, they argued that the Shibata's idea to fit the model to another independent realization is unrealistic since in practice we only have one data at hand. The common feature of these works is their asymptotic framework.
Meanwhile, there were several authors which study this question in non asymptotic regime. Goldenshluger and Zeevi (2001) in the non parametric framework, studied how well a Gaussian process admitting an AR(∞) representation can be approximated by a finite-order AR model.

In Baraud et al. (2001a) and Baraud et al. (2001b), they analyzed similar question, but a little bit different as observations arise from an auto-regressive model of order k. They proved an oracle inequality under several conditions, for instance the compactly supported base of the regression function. Moreover, they assume that the process is β-mixing which is usually admitted, but quite hard to verify in practice. For linear processes, the τ-mixing is more suitable since its coefficients can be easily computed (see Comte et al. (2008)) and be bounded by a function of the model parameter θ^* (see Doukhan and Wintenberger (2008)). In this work, we do not assume any mixing property of the process since the condition **A1** implies the τ-mixing property (see Doukhan and Wintenberger (2008)) and we will see that the decreasing rate of τ-mixing coefficients is bounded by the decreasing rate of the coefficients $\theta^* = (\theta_i^*)_{i\in\mathbb{N}}$.
Based on the above and following a model selection approach, our purpose in this work is to design adaptive penalties in such a way that the selected model mimic the oracle when observations arise form AR(∞) under mild conditions, including the existence of the all order moment of the noise, the decreasing rate of the coefficients of $(\theta_i^*)_{i\in\mathbb{N}}$ so that thanks to a result by Doukhan and Wintenberger (2008), the generating process has nice properties such as stationarity, τ-mixing.
The main contribution of this paper is to have proved that the excess risk of the selected estimator enjoys the best bias-variance trade-off over the considered collection.
The paper is organized as follows. The model selection approach along with preliminary results are described in Section 5.2. The main results are presented Section 5.3. Finally, Section 5.4 contains the proofs.

5.2 Model Selection Approach and Preliminary Results

5.2.1 Model Selection Approach

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

$$f(x_1, x_2, \ldots, x_{D_m}) = \sum_{i=1}^{D_m} \theta_i \, x_i, \tag{5.2.1}$$

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

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

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

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

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

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

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

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

Its empirical version minimizing the least-squares contrast is

$$\widehat{\theta}_m = \operatorname*{argmin}_{\theta \in \Theta_m} \gamma_n(\theta) \quad \text{where} \quad \gamma_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} (X_t - f_\theta^t)^2. \tag{5.2.2}$$

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

$$\ell(\widehat{\theta}, \theta^*) = \mathbb{E}\Big[\big\|F_{\widehat{\theta}} - F_{\theta^*}\big\|_n^2\Big] \tag{5.2.3}$$

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

Given that all the models which can be considered must have finite dimensions for fixed n, making all S_m wrong models, it is classical to let the dimension of competitive models grow with the number of observations. This will help reduce the excess loss and provide a better approximation of f_{θ^*}.

Let $\mathcal{M}_n$ a countable collection of hierarchical model S_m and K_n is the dimension of the largest model in $\mathcal{M}_n$ satisfying $|\mathcal{M}_n| \leq K_n < n$. We follow the classical approach of model selection which consists in minimizing the penalized LSE. Let pen: $\mathcal{M}_n \to \mathbb{R}^+$ be a penalty function, possibly data-dependent, and define

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{M}_n} \{C(m)\} \qquad \text{with} \qquad C(m) := \gamma_n\big(\widehat{\theta}_m\big) + \operatorname{pen}(S_m). \tag{5.2.4}$$

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

$$m^* \in \arg \inf_{m \in \mathcal{M}_n} \ell(\widehat{\theta}_m, \theta^*). \tag{5.2.5}$$

The oracle m^* is unachievable since it depends on θ^* and the distribution $P_{(X_1,\dots,X_n)}$ that are unknowns. However, we hope to select a model $\widehat{m}$ so that $\ell(\widehat{\theta}_{\widehat{m}}, \theta^*)$ is closest to $\ell(\widehat{\theta}_{m^*}, \theta^*)$.
The goal of this paper is to propose a data driven penalty in order to obtain an oracle inequality

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1 \inf_{m \in \mathcal{M}_n} \left\{ \ell(\widehat{\theta}_m, \theta^*) \right\} + \frac{C_2}{n} \tag{5.2.6}$$

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

$$\ell(\widehat{\theta}_{\widehat{m}}, \theta^*) \leq C_1' \inf_{m \in \mathcal{M}_n} \left\{ \ell(\theta_m^*, \theta^*) + \text{pen}(S_m) \right\} + \frac{C_2'}{n} \tag{5.2.7}$$

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

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

5.2.2 Notations

We will use the following norms:

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

5.2.3 Preliminary Results

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

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

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

where $\Lambda(E)$ is the set of 1-Lipschitz function, i.e. the functions f from $\left(E, \|.\|_E\right)$ to $\mathbb{R}$ such that $|f(x) - f(y)| \leq \|x - y\|_E$.
Using the definition of τ, we will measure the dependence of the strictly stationary sequence $(Z_t)_{t\in\mathbb{Z}}$ thanks to the coefficients defined as follows. For any $s \geq 0$, let introduce the norm $\|x - y\|_{\mathbb{R}^k} = (|x_1 - y_1| + \cdots + |x_k - y_k|)$ and setting $\mathcal{M}_i = \sigma(Z_t, t \leq i)$ and if $\mathbb{E}(|Z_1|) < \infty$, let

$$\tau_{Z,\infty}^{(p)}(s) \quad = \quad \sup_{l>0} \left\{ \max_{1 \leq k \leq l} \frac{1}{k} \sup \left\{ \tau^{(p)}\left(\mathcal{M}_i, (Z_{i_1}, \dots, Z_{i_k})\right) \ , i + s \leq i_1 < \cdots < i_k \right\} \right\}.$$

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

The next Proposition that is a consequence of Theorem 3.1 in Doukhan and Wintenberger (2008) gives a link between the τ-mixing coefficients of the process $(X_t)_{t\in\mathbb{Z}}$ and the coefficients θ^*_i of the model (5.1.3).

Proposition 5.1. *Assume **A1** holds and if $|\theta^*_t| = O(t^{-\gamma})$ with $\gamma > 1$, there exists a τ-weakly dependent stationary solution of* (5.1.1) *and a constant $C_\tau > 0$ such that for $r > 0$*

$$\tau^{(2)}_{X,\infty}(r) \le C_\tau \left(\frac{\log r}{r}\right)^{\gamma-1} \tag{5.2.8}$$

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

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

Therefore (5.2.8) is a straightforward application of Theorem 3.1 in Doukhan and Wintenberger (2008). □

As we are going to need independence for block of random variables, let denote for $t = 1,\dots,n$ the random vector $\vec{X}_t := (X_{t-1},\dots,X_{t-K_n})^\top$ One can see that the process $(\vec{X}_t)_{t\in\mathbb{Z}}$ is also mixing with $\tau^{(1)}_{\vec{X},\infty}$ upper bounded by $K_n\,\tau^{(1)}_{X,\infty}$ (see Lemma 5.1).
Now, we construct random variables approximating $\vec{X}_t$'s enjoying the independence by block property. Let s_n, q_n two integers such that $n = 2\,s_n\,q_n$. We are going to build $2\,s_n$ blocks of length q_n so that the even index blocks are independent and so the odd index blocks.
For $k = 0,\dots,s_n - 1$ let denote by

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

We recall a result of Lerasle et al. (2011) which is a consequence of the coupling in Dedecker and Prieur (2005).

Proposition 5.2. *Let $(X_t)_{t\in\mathbb{Z}}$ be the stationary mixing process process obtained in Proposition 5.1. Let also s_n, q_n, A_k, B_k defined as above for $k = 0,\dots,s_n-1$. There exist random vectors $A^*_k = \big(\vec{X}^*_{2kq_n+1},\dots,\vec{X}^*_{(2k+1)q_n}\big)$, $B^*_k = \big(\vec{X}^*_{(2k+1)q_n+1},\dots,\vec{X}^*_{(2k+2)q_n}\big)$ such that:*

1. *For $k = 0,\dots,s_n-1$, A^*_k has the same law as A_k, also B^*_k and B_k.*
2. *The random vectors $(A^*_k)_{0\le k\le s_n-1}$ are independent and so are the vectors $(B^*_k)_{0\le k\le s_n-1}$.*
3. $$\big\|A_k - A^*_k\big\|_1 \le q_n\,K_n\,\tau^{(1)}_{X,\infty}(q_n) \qquad \text{and} \quad \big\|B_k - B^*_k\big\|_1 \le q_n\,K_n\,\tau^{(1)}_{X,\infty}(q_n).$$

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

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

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

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

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

where the Inequality follows from Hölder's Inequality.

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

A3: For any $f_\theta \in S_m$, $< \alpha, \partial_\theta f_\theta >= 0$ *a.s.* $\implies$ $\alpha = 0$
This condition means that the columns of the matrix $\mathbf{M}_m$ are linearly independents.

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

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

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

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

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

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

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

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

and then it is simple to see that

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

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

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

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

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

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

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

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

Proposition 5.3. *Under **A1** with $|\theta_t^*| = O(t^{-\gamma})$ where $\gamma \geq 2$, we have for any $m \in \mathcal{M}_n$*

$$\big\|\Sigma_m\big\|_{\rm op} \leq \pi^{-1}\sum_{i=0}^{\infty}\big|\mathbb{E}[X_0\,X_i]\big| < \infty. \tag{5.2.9}$$

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

$$\big\|\Sigma_m^{-1}\big\|_{\rm op} \leq 1/a. \tag{5.2.10}$$

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

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

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

$$\mathbf{A5}: \qquad \Big(\frac{\log q_n}{q_n}\Big)^{\gamma-1} \leq \frac{A}{n}, \tag{5.2.11}$$

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

$$\mathbf{A6}: \quad \frac{s_n}{2} \min\left\{\left(\frac{1}{2^7\,\sigma_0^2\,K_n}\right)^2, \frac{1}{2^8\,\sigma_0^2\,K_n}\right\} \geq 3\log n, \tag{5.2.12}$$

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

Proposition 5.4. *Under assumptions **A1**, **A6** and if $|\theta_t^*| = O(t^{-\gamma})$ with $\gamma \geq 8$, there exists a constant C such that*

$$\mathbb{P}(\Omega_n^c) \leq \frac{C}{n^3}. \tag{5.2.13}$$

5.3 Bias-Variance Result

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

Theorem 5.1. *Let consider observations $(X_1, \ldots, X_n)$ arising from a solution of the process (5.1.1) satisfying **A1** with $|\theta_t^*| = O(t^{-\gamma})$ where $\gamma \geq 8$ and also verifying **A2** and **A4**. Let $\mathcal{M}_n$ be some countable family of AR models satisfying **A3** and **A5-A6**. For any constant $x > 0$, let a penalty function pen: $\mathcal{M}_n \to \mathbb{R}^+$ such that*

$$\mathrm{pen}(S_m) \geq 8\,x^3\sigma^2\,\frac{D_m}{n}. \tag{5.3.1}$$

Then, the LSE $\widehat{\theta}_{\widehat{m}}$ with $\widehat{m}$ given in (5.2.4), satisfies

$$\mathbb{E}\Big[\big\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\big\|_n^2 \mathbb{1}_{\Omega_n}\Big] \leq C_1(x) \inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\Big[\big\|F_{\theta^*} - F_{\theta_m^*}\big\|_n^2\Big] + 2\,\mathrm{pen}(S_m)\right\} + \frac{x(x+2)}{x-2}\,\frac{C_2}{n} \tag{5.3.2}$$

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

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

5.4 Proofs

5.4.1 Proof of Theorem 5.1

Proof. We follow the scheme of the proof of Baraud et al. (2001b). Let fix $m \in \mathcal{M}_n$. From the definition (5.2.4), we have

$$\gamma_n\big(\widehat{\theta}_{\widehat{m}}\big) + \mathrm{pen}(S_{\widehat{m}}) \leq \gamma_n\big(\widehat{\theta}_m\big) + \mathrm{pen}(S_m) \leq \gamma_n(\theta_m^*) + \mathrm{pen}(S_m). \tag{5.4.1}$$

Since,

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

(5.4.1) yields to

$$\left\|F_{\widehat{\theta}_{\widehat{m}}} - F_{\theta^*}\right\|_n^2 \le \left\|F_{\theta^*_m} - F_{\theta^*}\right\|_n^2 + \frac{2\,\sigma}{n}\sum_{t=1}^{n}\xi_t(f^t_{\widehat{\theta}_{\widehat{m}}} - f^t_{\theta^*_m}) + \mathrm{pen}(S_m) - \mathrm{pen}(S_{\widehat{m}}). \quad (5.4.2)$$

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

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

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

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

Moreover, on the set Ω_n, it holds

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

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

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

Therefore, from (5.4.2), it holds on Ω_n

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

where $D(S_{\widehat{m}}) = dim(S_m + S_{\widehat{m}}) \leq D_m + D_{\widehat{m}}$. Hence using the condition on the penalty (5.3.1),

$$\Big(1-2\,\frac{(1+y)}{x}\Big)\,\big\|F_{\widehat{\theta}_{\widehat{m}}}-F_{\theta^*}\big\|_n^2 \leq \Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\,\big\|F_{\theta^*}-F_{\theta^*_m}\big\|_n^2+2\,\mathrm{pen}(S_m)+x\,\sigma^2\,V_{\widehat{m}} \tag{5.4.3}$$

with

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

where $\nu_n(g_\theta) := n^{-1}\sum_{t=1}^n \xi_t\, g_\theta^t$.
The proof will be established after controling the expectation of $V_{\widehat{m}}$ which involves the supremum of an empirical process.

Now we leverage the mixing property in order to apply Talagrand's Inequality (Theorem 5.2) to tackle $\mathbb{E}[V_{\widehat{m}}]$.

We have

$$V_{\widehat{m}} = \left[\sup_{g_\theta\in B(\widehat{m},\mu)} \big(\nu_n(g_\theta)\big)^2 - 8\frac{x^2}{n}\,D(S_{\widehat{m}})\right]_+ \leq 2\sup_{g_\theta\in B(\widehat{m},\mu)}\Big(\nu_n(g_\theta)-\nu_n^*(g_\theta)\Big)^2 + V^*_{\widehat{m}} \tag{5.4.4}$$

where

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

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

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

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

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

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

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

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

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

where the last inequality follows from Proposition 5.1. Thus,

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

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

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

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

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

with

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

where

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

Now let remark that $\nu_{n,1}^*(g_\theta)$ and $\nu_{n,2}^*(g_\theta)$ are both sum of s_n independent random variables by virtue of Proposition 5.2. Hence,

$$\begin{aligned}V_{\widehat{m}}^* \leq &\Big(\sup_{g_\theta \in B(\widehat{m},\mu)} 4\left(\nu_{n,1}^*(g_\theta)\right)^2 - 4\, x^2\, n^{-1}\, D(S_{\widehat{m}})\Big)_+ \\ &+ \Big(\sup_{g_\theta \in B(\widehat{m},\mu)} 4\left(\nu_{n,2}^*(g_\theta)\right)^2 - 4\, x^2\, n^{-1}\, D(S_{\widehat{m}})\Big)_+.\end{aligned}$$

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

Bounding $\mathbb{E}_1^*$

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

$$\xi_t = \eta_t + \epsilon_t, \quad \eta_t = \xi_t \mathbb{I}_{|\xi_t| \leq k_n},$$

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

$$\nu_{n,1}^*(g_\theta) = \upsilon_{n,1}^*(g_\theta) + \upsilon_{n,2}^*(g_\theta), \quad \text{where}$$

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

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

Thus,

$$\begin{aligned}
\mathbb{E}_1^* &\leq 8\,\mathbb{E}\left[\Big(\sup_{g_\theta \in B(\widehat{m},\mu)} \big(\upsilon_{n,1}^*(g_\theta)\big)^2 - 0.5\,x^2\,n^{-1}\,D(S_{\widehat{m}})\Big)_+\right] + 2\,\mathbb{E}\left[\sup_{g_\theta \in B(\widehat{m},\mu)} \big(\upsilon_{n,2}^*(g_\theta)\big)^2\right] \\
&\leq 8 \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[\Big(\sup_{g_\theta \in B(m',\mu)} \big(\upsilon_{n,1}^*(g_\theta)\big)^2 - 0.5\,x^2\,n^{-1}\,D(S_{m'})\Big)_+\right] \qquad (5.4.5) \\
&\qquad + 2\,\mathbb{E}\left[\sup_{g_\theta \in B(\widehat{m},\mu)} \big(\upsilon_{n,2}^*(g_\theta)\big)^2\right]. \qquad (5.4.6)
\end{aligned}$$

We start by bounding the term in (5.4.5). Let $m' \in \mathcal{M}_n$. In order to apply Theorem 5.2, one has to find M, H and v such that

$$\sup_{g_\theta \in B(m',\mu)} \big|\upsilon_{n,1,k}^*(g_\theta)\big| \leq M, \quad \mathbb{E}\Big[\sup_{g_\theta \in B(m',\mu)} \big|\upsilon_{n,1}(g_\theta)\big|^2\Big] \leq H^2,$$

$$\text{and} \quad \sup_{g \in B(m',\mu)} \mathrm{Var}\left(\upsilon_{n,1,k}^*(g_\theta)\right)\big) \leq v.$$

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

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

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

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

As a result, we have

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

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

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

Moreover,

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

since $D(S_{m'}) \geq 1$.
• Lastly, as $\mathrm{Var}\,[X] \leq E[X^2]$, it follows from the previous series of equations

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

As a consequence from Theorem 5.2 and taking $\alpha = \frac{1}{2}(\frac{x^2}{2\Phi_0^2\,k_n^2} - 1) > 0$, we have

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

Hence there exists a constant K' such that

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\Big(\sup_{g_\theta \in B(m',\mu)}\big(\upsilon^*_{n,1}(g_\theta)\big)^2 - 0.5\,x^2\,n^{-1}\,D(S_{m'})\Big)_+\right] \leq \frac{K'}{n}. \tag{5.4.7}$$

• Now, let us upper bound the term in (5.4.6). For any $m' \in \mathcal{M}_n$ and any $g_\theta \in B(m',\mu)$, we have

$$g_\theta(\vec{X}_t) \;=\; \sum_{i=1}^{D(S_{m'})}\theta_i\,X_{t-i} \leq \sup_{t-D(S_{m'})\leq i<t}|X_i|\,\Big(\sum_{i=1}^{D(S_{m'})}|\theta_i|\Big) < \sup_{t-D(S_{m'})\leq i<t}|X_i|.$$

Therefore,

$$\begin{aligned}
\upsilon^*_{n,2,k}(g_\theta) &= \frac{1}{2\,q_n}\sum_{i=1}^{q_n}\xi_{2kq_n+i}\,g_\theta(\vec{X}^*_{2kq_n+i}) \\
&< \frac{1}{2\,q_n}\sum_{i=1}^{q_n}|\xi_{2kq_n+i}|\sup_{2kq_n+i-K_n\leq t<2kq_n+i}|X^*_t| := Y^*_k,
\end{aligned} \tag{5.4.8}$$

so that

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

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

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

Moreover,

$$\begin{aligned}\mathbb{E}\big[Y^{*2}_0\big] &= \frac{1}{4\,q_n^2}\sum_{i,j=1}^{q_n}\mathbb{E}\Big[|\xi_i|\sup_{i-K_n\le t<i}|X^*_t|\,|\xi_j|\sup_{j-K_n\le t<j}|X^*_t|\Big]\\ &= \frac{1}{4\,q_n^2}\sum_{i=1}^{q_n}\mathbb{E}\Big[\Big(\sup_{i-K_n\le t<i}|X^*_t|\Big)^2\Big]\\ &= \frac{\mu_2}{4\,q_n},\end{aligned} \tag{5.4.9}$$

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

$$\mathbb{E}\Big[\sup_{g_\theta\in B(\widehat{m},\mu)} \big|\nu^*_{n,1}(g_\theta)\big|^2\Big] < \frac{\mu_2}{4\,s_n\,q_n} = \frac{\mu_2}{2\,n}. \tag{5.4.10}$$

Inequality (5.4.7) along with (5.4.10) yields to

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

We conclude that there exists $K>0$

$$\mathbb{E}[V_{\widehat{m}}] \le \frac{K}{n}. \tag{5.4.11}$$

Returning to (5.4.3), and taking expectation on both sides, it then follows

$$\Big(1-2\,\frac{(1+y)}{x}\Big)\,\mathbb{E}\Big[\big\|f^t_{\widehat{\theta}_{\widehat{m}}}-f^t_{\theta^*}\big\|_n^2\Big] \le \Big(1+2\,\frac{(1+y^{-1})}{x}\Big)\,\mathbb{E}\Big[\big\|f^t_{\theta^*}-f^t_{\theta^*_m}\big\|_n^2\Big]+2\,\mathrm{pen}(S_m)+x\,\frac{K}{n}. \tag{5.4.12}$$

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

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

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

5.4.2 Proof of Proposition 5.4

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

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

where

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

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

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

Moreover,

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

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

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

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

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

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

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

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

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

Indeed,

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

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

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

Using Lemma 5.3 with $u = 1/4$ and by virtue of **A6**, it follows

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

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

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

Thus, from Markov's Inequality,

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

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

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

Hence using Proposition 5.1, it follows

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

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

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

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

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

As a consequence,

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

□

5.4.3 Proof of Proposition 5.3

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

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

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

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

But, from Lemma 5.2 and since $\gamma \geq 2$, it follows

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

Given that Σ_m is symmetric, it follows

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

which concludes the proof of (5.2.9).

Now we end by the proof of (5.2.10). Reasoning as above, and by virtue of **A4**, one can show that

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

which yields to

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

so that (5.2.10) is established.

□

5.4.4 Technical Lemmas

Lemma 5.1. *Assume **A1** holds and (X_t) the mixing stationary solution of* (5.1.1)*. Then, the process $(\vec{X}_t)$ is mixing and*

$$\tau^{(1)}_{\vec{X},\infty}(r) \leq K_n\, \tau^{(1)}_{X,\infty}(r-1). \tag{5.4.13}$$

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

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

As a result,

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

This fact along with the definition of $\tau^{(1)}_{\vec{X},\infty}(r)$ leads to (5.4.13).

□

Lemma 5.2. *Under **A1** with $|\theta^*_t| = O(t^{-\gamma})$ where $\gamma > 1$, we have*

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

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

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

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

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

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

Lemma 5.3. *Under assumptions **A2**, it holds for any model $m \in \mathcal{M}_n$, and for all $u > 0$*

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

Proof. One can write for a matrix A

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

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

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

But,

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

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

Moreover, we can write

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

Therefore,

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

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

$\{\mathbf{Y}_{1,k}\}$ and $\{\mathbf{Y}_{2,k}\}$ are independent random vectors by virtue of Proposition 5.2. Now, let us show that $\mathbf{Y}_{i,k}$ are sub-exponentials. For λ such that $|\lambda| < \frac{1}{16\,D_m\sigma_0^2}$, and denoting

$w_i = Y^2_{2kq_n+i} - \mathbb{E}[Y_1^2]$, we have

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

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

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

so that

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

□

Lemma 5.4. *Assume **A3** holds, then $\widehat{\Sigma}_m$ is a.e. invertible. Also, Σ_m is invertible.*

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

Moreover, $\Sigma_m = \mathbb{E}\big[\widehat{\Sigma}_m\big] = \mathbb{E}\big[Z_0^m(Z_0^m)^\top\big]$ with $Z_0^m = (X_{-1},\ldots,X_{-D_m})^\top$. Let $\mathbf{u}\in\mathbb{R}^{D_m}$, it follows $\mathbf{u}^\top\Sigma_m\mathbf{u} = \mathbb{E}\big[((Z_0^m)^\top\mathbf{u})^2\big] \geq 0$. Let show that whenever the equality holds ($\mathbf{u}^\top\Sigma_m = 0$), $\mathbf{u} = 0$.
Since $((Z_0^m)^\top\mathbf{u})^2 \geq 0$, its expectation vanishes if and only if $(Z_0^m)^\top\mathbf{u} = 0$ a.e. which yields to $\mathbf{u} = 0$ by **A3**. Hence, Σ_m is positive definite and then invertible. □

5.5 Theoretical Tools

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

Theorem 5.2. *Let $Y_1,\ldots,Y_n$ be independent random variables and let $\mathcal{F}$ be a countable class of uniformly bounded measurable functions. Then for all $\alpha > 0$,*

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

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

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

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

6

Conclusion Générale et Perspectives

Dans cette thèse, il était question pour nous de proposer et d'étudier sur les plans théorique et numérique des critères de sélection de modèles pour les séries affines causales.

Dans un premier temps, nous avons montré qu'il était indispensable de faire dépendre la pénalité du critère de sélection de la vitesse de décroissance des coefficients de Lipschitz du processus afin d'obtenir un critère consistant en probabilité. Nous avons également proposé un test d'adéquation du modèle sélectionné basé sur l'autocorrélation du carré des résidus du modèle. Les simulations numériques ont montré des résultats assez satisfaisants.

Deuxièmement, nous avons proposé une généralisation du critère de Hannan et Quinn à la classe des séries affines causales. Cette généralisation fournit une nouvelle constante multiplicative (dans la pénalité) connue pour les modèles classiques (ARMA, GARCH ou APARCH) et data-driven calibrable pour les modèles complexes tels que les ARMA-GARCH. Là également, quelques études de simulation ont attesté de la bonne qualité des critères obtenus.

Dans un troisième temps, nous nous sommes intéressés à la construction des critères asymptotiquement efficients. Nous avons généralisé le critère AIC d'Akaike en s'appuyant sur la pénalité dite idéale. Le comportement asymptotique de cette pénalité idéale nous a suggéré un terme de pénalité qui vaut exactement $2\,D_m$ comme dans l'AIC pour des modèles assez simples, et pour des modèles complexes, nous avons donné une formule moins explicite mais qui peut être calibrée au moyen des données. Dans cette troisième partie, nous avons également à la suite de Schwartz, dérivé le critère BIC basé sur la maximisation de la probabilité a posteriori de choisir le vrai modèle.

Enfin, nous nous sommes restreints à l'étude non asymptotique d'un processus particulier de la classe des modèles affines causaux. Un estimateur des moindres carrés pénalisé est construit à partir d'un critère de sélection adaptatif et la sélection est opérée parmi une collection de modèles linéaires. Nous avons montré que l'estimateur final est presque aussi performant que le meilleur sur la collection considérée, *i.e.* qu'il réalise, à une constante près, le compromis biais-variance. La pénalité obtenue généralise celle de Mallows

et dépend d'une constante que l'on pourrait estimer avec des algorithmes de calibration adaptative.

Au sortir de cette thèse, il est important de souligner que nos travaux ouvrent plusieurs perspectives de recherche.

1/ A la suite du Chapitre 4, il serait intéressant d'obtenir l'inégalité d'oracle (5.2.6) cette fois ci en considérant que les observations proviennent du modèle AR(∞) suivant:

$$X_t = \sum_{k=1}^{+\infty} \theta_k^* \, X_{t-k} + \sigma \, \xi_t \quad \text{pour tout } t \in \mathbb{Z}. \tag{6.0.1}$$

où $(\xi_t)_{t\in\mathbb{Z}}$ est un bruit blan **faible**. La tâche devient beaucoup plus difficile car l'hypothèse d'indépendance du bruit est fondamentale pour garantir l'existence d'une solution mélangeante, propriété nécessaire pour utiliser des inégalités exponentielles usuelles. L'on pourrait ainsi dans un premier temps, supposer que le processus $(X_t)_{t\in\mathbb{Z}}$ est mélangeant. Cela pourrait permettre de traiter les GARCH puisque l'on peut toujours les écrire comme des ARMA à bruit blanc faible.

2/ Une extension logique du Chapitre 4 serait de considérer que les observations proviennent d'un modèle fonctionnel autoregréssif

$$X_t = f^*(X_{t-1}, X_{t-2}, \ldots) + \sigma \, \xi_t \quad \text{pour tout } t \in \mathbb{Z} \tag{6.0.2}$$

avec bruit blanc fort mais où la fonction f^* n'est plus nécessairement linéaire mais vérifie l'hypothèse de contraction

$$\big|f^*(x) - f^*(y)\big| \leq \sum_{i=1}^{\infty} \alpha_i \, |x_i - y_i|$$

pour tout $x, y \in \mathbb{R}^\infty$ avec $\sum_{i=1}^{\infty} \alpha_i < 1$.

Un modèle candidat S_m à considérer pourrait être dans ce cas, un D_m sous-espace linéaire de $\mathbb{L}^2(\mathbb{R}^{D_m})$. Si $\mathcal{M}_n$ désigne la famille de modèles candidats, l'on définit l'estimateur pour tout $S_m \in \mathcal{M}_n$ comme

$$\widehat{f}_m = \underset{f\in S_m}{\operatorname{argmin}} \, \gamma_n(f) \quad \text{avec} \quad \gamma_n(f) = \frac{1}{n} \sum_{t=1}^{n} \big(X_t - f(X_{t-1}, X_{t-2}, \ldots, X_{t-D_m})\big)^2.$$

A travers une procédure de pénalisation, l'objectif serait de trouver la bonne fonction de pénalité pen de sorte qu'on ait

$$\ell(\widehat{f}_{\widehat{m}}, f^*) \leq C_1 \inf_{m\in\mathcal{M}_n} \big\{\ell(\widehat{f}_m, f^*)\big\} + C_2 \, n^{-1} \tag{6.0.3}$$

avec

$$\widehat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \, \{C(m)\} \qquad \text{avec} \qquad C(m) := \gamma_n\big(\widehat{\theta}_m\big) + \operatorname{pen}(S_m), \tag{6.0.4}$$

$\ell(\widehat{f}_m, f^*) = \mathbb{E}\big[(X_t - \widehat{f}_m(X_{t-1}, X_{t-2}, \ldots, X_{t-D_m}))^2\big] - \sigma^2$ et la constante C_1 proche de 1.

3/ De plus en plus, l'hypothèse de stationnarité est remise en question et l'intérêt pour des séries non stationnaires surtout localement stationnaires est croissant ces dernières années. Bardet et al. (2020a) ont prouvé des résultats généraux (consistance, TLC) pour des processus localement stationnaires. Il serait intéressant de s'appuyer sur ces résultats pour faire de la sélection de modèles sur ces processus.

Bibliography

H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

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

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

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

O. Arkoun, J.-Y. Brua, and S. Pergamenshchikov. Sequential model selection method for nonparametric autoregression. *Sequential Anal.*, 38(4):437–460, 2019.

S. Arlot. *Rééchantillonnage et Sélection de modeles*. PhD thesis, Université Paris Sud-Paris XI, 2007.

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

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

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

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

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

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(1):452–479, 2017.

J.-M. Bardet, P. Doukhan, and O. Wintenberger. Contrast estimation of general locally stationary processes using coupling. *Preprint arXiv:2005.07397*, 2020a.

J.-M. Bardet, K. Kamila, and W. Kengne. Consistent model selection criteria and goodness-of-fit test for common time series models. *Electronic Journal of Statistics*, 14(1):2009–2052, 2020b.

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

E. Beale, M. Kendall, and D. Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357–366, 1967.

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

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

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

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

H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

P. Brockwell and R. Davis. *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media, 1991.

J. Cavanaugh. Unifying the derivations for the akaike and corrected akaike information criteria. *Statistics & Probability Letters*, 33(2):201–208, 1997.

A. Celisse. *Model selection via cross-validation in density estimation, regression, and change-points detection*. PhD thesis, Université Paris Sud-Paris XI, 2008.

A. Celisse et al. Optimal cross-validation in density estimation with the l^{2}-loss. *Annals of Statistics*, 42(5):1879–1910, 2014.

C.-F. Chen. On asymptotic normality of limiting density functions with bayesian omplications. *J. R. Statist. Soc. B*, 47:540–546, 1985.

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

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

I. Csiszár and P. Shields. Consistency of the bic order estimator. *Electronic research announcements of the American mathematical society*, 5(17):123–127, 1999.

I. Csiszár, P. C. Shields, et al. The consistency of the bic markov order estimator. *The Annals of Statistics*, 28(6):1601–1619, 2000.

J. Dedecker and P. Doukhan. A new covariance inequality and applications. *Stochastic Processes and Applications*, 106:63–80, 2003.

J. Dedecker and C. Prieur. Coupling for $\tau-$Dependent Sequences and Applications. *Journal of Theoretical Probability*, 17:861–885, 2004.

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

J. Dedecker, P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur. *Weak dependence: With Examples and Applications*. Lecture Notes in Statistics **190**, Springer-Verlag, New York, 2007.

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

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

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

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

D. F. Findley and C.-Z. Wei. Aic, overfitting principles, and the boundedness of moments of inverse matrices for vector autotregressions and related models. *Journal of Multivariate Analysis*, 83(2):415–450, 2002.

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

C. Francq and J.-M. Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2010.

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

A. Garivier. Consistency of the unlimited bic context tree estimator. *IEEE Transactions on Information theory*, 52(10):4630–4635, 2006.

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

U. S. government. Particulate matter (pm) pollution. `http://www.epa.gov/pm-pollution/particulate-matter-pm-basics`, 2017.

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

E. J. Hannan and M. Deistler. *The statistical theory of linear systems*. SIAM, 2012.

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, 1979.

R. R. Hocking and R. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

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

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

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(2):1061–1087, 2019a.

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

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

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

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

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

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

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

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

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

R. L. Kashyap. Optimal choice of ar and ma parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):99–104, 1982.

R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on laplace's method. *Essays in Honor of George Barnard, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland*, pages 473–188, 1990.

W. Kengne. Strongly consistent model selection for general causal time series. *Statistics & Probability Letters*, page 109000, 2020.

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

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

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

É. Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4):717–736, 2005.

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

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

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

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

K.-C. Li. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

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

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

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

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

J. Lv and J. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B*, 76(1):141–167, 2014.

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

P. Massart. *Concentration inequalities and model selection.* Springer, 2007.

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

F. Papangelou. On a distributional bound arising in autoregressive model fitting. *Journal of applied probability*, 31(2):401–408, 1994.

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

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

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

S. L. Sclove. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343, 1987.

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(2):507–524, 2017.

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

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

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

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

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

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

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

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

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

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25, 1982.

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

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.