

Estimating linear representations of nonlinear processes

Christian Francq^a, Jean-Michel Zakoïan^{a,b,*}

^a *Université de Lille I, Département de mathématiques, Lab. de Probabilités et Statistique,
59655 Villeneuve d'Ascq Cedex, France*

^b *CREST, Batiment Malakoff 2-Timbre J320, 92245 Malakoff Cedex, France*

Received 11 March 1996; received in revised form 16 December 1996

Abstract

This paper considers the problem of estimating an autoregressive-moving average (ARMA) model when only ergodic and mixing assumptions can be made. The estimation procedure is based on the minimization of a sum of squared deviations about linear conditional expectations. It is shown that the estimator is strongly consistent and asymptotically normal. The results can be used to estimate weak linear representations of some nonlinear processes. Several examples of such linear representations are provided. Other potential areas of applications are inference for noncausal ARMA, aggregation and marginalization of linear processes. A numerical study is also presented. It appears that standard identification routines based on strong hypothesis on the innovation of ARMA models can be seriously misleading when these assumptions do not hold. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: ARMA models; Nonlinear models; Least-squares estimator; Consistency; Asymptotic normality

1. Introduction

In the last 15 years, the time-series literature has been characterized by a growing interest in nonlinear models. A wide variety of formulations have been proposed and extensively studied (see e.g. Priestley (1988) or Tong (1990) for reviews on some recent work on nonlinear time-series analysis). Their usefulness in applied research has been demonstrated in numerous cases. However, they seem often complex to handle and linear models remain the most widely used by the practitioners. Ready-made packages are available, however, they rely on strong assumptions on the noise processes (such as independence or martingale difference). These assumptions are typically not satisfied for linear representations of nonlinear processes. It is, therefore, interesting to investigate the features of the Box and Jenkins methodology in cases where the standard strong hypothesis on the linear innovation do not hold. This paper is a first step

* Corresponding address. CREST, Batiment Malakoff 2-Timbre J320, 92245 Malakoff Cedex, France.
E-mail: zakoïan@ensae.fr.

in this direction since we focus on the estimation problem of ARMA representations of univariate nonlinear processes.

A well-known property of weak ARMA processes is that they can be viewed as L^2 approximations of the Wold decomposition of regular stationary processes. Besides, several important classes of nonlinear models have been found to admit *exact* weak ARMA representations. One may ask why one needs to estimate the latter. There are at least three reasons. One is that they provide invaluable assistance in forecasting. Although not optimal in the L^2 sense, the linear forecasts are easily obtained and do not rely on a complete specification of the dynamics. Second, a linear representation can be a helpful tool for model selection. Given the range of alternative nonlinear models, the selection of a particular class is often a difficult task and the compatibility of one particular candidate with the ARMA representation can be worthwhile noting. The third reason concerns model identification. Several nonlinear models (such as the GARCH) include a linear part, completed by a modelling of the dynamics of the linear innovation. In such cases the linear representation is typically used to identify the nonlinear part in a two-steps procedure.

Another potential area of application of the results stated in the present paper concerns inference for noncausal ARMA processes, with non-Gaussian i.i.d. noise. These processes are well known to admit causal ARMA representations, however, such representations are generally weak ones. We are thankful to a referee who pointed out this application which we believe to be promising.

Finally, other potentially fruitful applications can be found within the class of strong linear processes. Transformations commonly used in applied research, such as aggregation, marginalization of vector time series, are well known to preserve linearity. However, as will be seen later on, such transformations fail to preserve *strong* linearity. Therefore, applying standard time series techniques to data obtained from transformed strong linear processes is likely to be misleading.

The estimation procedure developed here is based on the minimization of a sum of squared deviations about conditional *linear* expectations. This is closely related to the “conditional least-squares” method considered by Klimko and Nelson (1978), with strong innovations replaced by linear ones. The interpretation of linear conditional expectation as an orthogonal projection on the Hilbert space generated by the linear past of the process motivates the method, which we can call “conditional linear least-squares”. Although considerable attention has been paid to the asymptotic properties of various estimators in ARMA models with strong innovations, little is known when the martingale difference assumption is relaxed. To our knowledge, the most general treatment is due to Dunsmuir and Hannan (1976) who proved consistency of estimators derived from a Gaussian likelihood (although Gaussianity was not assumed) in a vector framework, under weak assumptions on the noise process and based on a spectral analysis (see also Hannan, 1975). They also obtained asymptotic normality under a martingale difference assumption. In this paper we show that a mixing property (in addition to the existence of moments) is sufficient to obtain a consistent and asymptotically normally distributed estimator for the parameters of a weak ARMA representation.

The paper is organized as follows. In Section 2 we provide some examples of nonlinear processes admitting a linear representation which we are able to exhibit. Several other potential areas of applications derived from linear processes are also described. The asymptotic results are presented in Sections 3 and 4 while the proofs and several lemmas are developed in Section 5. In Section 6, the asymptotic variance matrix of our estimator is derived for particular examples and its estimation is discussed. Finally, the results of a simulation study and an application to a real data set are presented.

2. Weak ARMA representations

Some models usually encountered in the nonlinear time series literature are based on a (weak) linear representation of the observable process. This is the case for the GARCH-type models introduced by Engle (1982), in which the linear innovation is modelled as a martingale difference with a specific form for the conditional variance. The approach can be extended to a more general framework (see Francq and Zakoian, 1995). The following are some less straightforward examples of nonlinear processes admitting ARMA representations.

2.1. ARMA representations of nonlinear processes

2.1.1. A process with a deterministic dynamics

Consider the following process, due to Moran and presented by Whittle (1963): let X_0 be a random variable uniformly distributed in $[0, 1]$, and, for $t \geq 1$, let X_t be the fractional part of $2X_{t-1}$. Easy calculations show that (X_t) admits the AR(1) representation: $X_t = \frac{1}{4} + \frac{1}{2}X_{t-1} + \varepsilon_t$, where (ε_t) is a weak white noise (i.e. a zero mean and uncorrelated second-order stationary sequence).

2.1.2. Bilinear processes

Initially studied by Granger and Anderson (1978) and Subba Rao (1978), bilinear models have been found to be useful in many areas. Pham (1985) has shown that a wide variety of bilinear processes admit ARMA representations. By way of illustration consider the following bilinear equation

$$\forall t \in \mathbb{Z}, \quad X_t = \eta_t + bX_{t-1}\eta_{t-2}, \quad (1)$$

where (η_t) is an i.i.d. $\mathcal{N}(0, 1)$ sequence. If $b^2 < 1$ then (1) admits a unique nonanticipative (i.e. measurable with respect to the σ -field generated by $\{\eta_u: u \leq t\}$) stationary solution (see e.g. Guégan, 1988). Straightforward algebra shows that this stationary solution (X_t) is also a MA(3):

$$\forall t \in \mathbb{Z}, \quad X_t = \varepsilon_t + c\varepsilon_{t-3}, \quad (2)$$

where (ε_t) is a weak white noise.

2.1.3. Switching-regime models

We consider a simple example of switching-regime Markov model (see, e.g., Hamilton (1994) for more details). Let (Δ_t) be a stationary, irreducible and aperiodic Markov chain with state space $\{0, 1\}$. The stationary distribution is defined by

$$\pi(0) = P(\Delta_t = 0) \quad \text{and} \quad \pi(1) = 1 - \pi(0) = P(\Delta_t = 1).$$

Let (η_t) be a sequence of i.i.d. centred variables with unit variance, which is supposed to be independent of (Δ_t) . Let (X_t) be the stationary process defined by

$$\forall t \in \mathbb{Z}, \quad X_t = \eta_t + (a + (b - a)\Delta_t)\eta_{t-1}. \tag{3}$$

Let $\gamma(h) = \text{Cov}(X_t, X_{t-h})$. We have

$$\begin{aligned} E(X_t) &= 0, \quad \gamma(0) = 1 + a^2\pi(0) + b^2\pi(1), \\ \gamma(1) &= a\pi(0) + b\pi(1) \quad \text{and} \quad \gamma(h) = 0, \quad \forall h > 1. \end{aligned}$$

Therefore, when $a\pi(0) + b\pi(1) \neq 0$, (X_t) admits the following MA(1) representation:

$$\forall t \in \mathbb{Z}, \quad X_t = \varepsilon_t + c\varepsilon_{t-1}, \tag{4}$$

where c is a constant depending on a, b and $\pi(0)$: $c = \gamma(0)/2\gamma(1) - \sqrt{(\gamma(0)^2/4\gamma(1)^2) - 1}$, and where (ε_t) is a weak white noise with variance equal to $\gamma(1)/c$.

2.1.4. Threshold models

The class of threshold autoregressive (TAR) processes has been introduced by Tong and Lim (1980). In general it seems difficult to exhibit, or even prove, the existence of ARMA representations for the TAR processes. A particular case of threshold process, introduced by Gouriéroux and Monfort (1992), is given by

$$Y_t = \sum_{j=1}^J \alpha_j I_{A_j}(Y_{t-1}) + \sum_{j=1}^J \beta_j I_{A_j}(Y_{t-1}) u_t,$$

where $(A_j, j = 1, \dots, J)$ is a partition of \mathbb{R}^d , the α_j 's are d -dimensional vectors, the β_j 's are positive-definite matrices and $(u_t)_{t \in \mathbb{Z}}$ is a sequence of i.i.d. random vectors with zero mean and covariance matrix equal to identity. Under an assumption on the distribution of Y_0 , (Y_t) can be shown to be strictly stationary and ergodic, and it admits an ARMA($J - 1, J - 1$) representation which can be obtained in closed form.

2.2. Weak representations derived from strong linear processes

The methods developed in this paper are also useful to deal with incompletely observed strong linear processes. There are at least three situations where an incomplete data problem can arise. First, the observations can be obtained with a time unit larger than that of the underlying process. Second, the observed process can result from the

linear combination of several independent unobservable processes. Finally, there are situations where only one or several components of a multivariate time series can be observable.

2.2.1. Temporal aggregation

In a large number of economic applications, some observations are missing at certain frequencies. A question of particular importance is whether an ARMA process at one frequency, say daily, is consistent with some ARMA process at another frequency, say weekly. The so-called temporal-aggregation properties of ARMA models have been studied extensively in the econometric literature (see, e.g., Amemiya and Wu, 1972; Palm and Nijman, 1984; Nijman and Palm, 1990).

Consider an ARMA(p, q) process $(X_t)_{t \in \mathbb{Z}}$ and define $(\tilde{X}_t)_{t \in \mathbb{Z}} = (X_{mt})_{t \in \mathbb{Z}}$, where m is any positive integer. Then $(\tilde{X}_t)_{t \in \mathbb{Z}}$ follows an ARMA($p, p + [(q - p)/m]$). However, the ARMA model is in general a weak one, i.e. the innovation process is not i.i.d., nor even a martingale difference. It is easily seen, for instance, by considering the following MA(2) process:

$$\forall t \in \mathbb{Z}, \quad X_t = \eta_t - \theta_1 \eta_{t-1} - \theta_2 \eta_{t-2},$$

where $\theta_1 \theta_2 \neq 0$ and (η_t) is an i.i.d. white noise with $E\eta_t^3 \neq 0$. Then $\tilde{X}_t = X_{2t}$ is solution of an MA(1) equation of the form: $\tilde{X}_t = \varepsilon_t - \theta_0 \varepsilon_{t-1}$, where $|\theta_0| < 1$. By inverting the MA(1) equation, one can obtain ε_t as a linear combination of η_{2t} and its past values. Hence, it is easily shown that, for instance,

$$E(\varepsilon_t \varepsilon_{t-1}^2) = E\eta_t^3 \left[(\theta_0 - \theta_2) + \frac{\theta_0}{1 - \theta_0^3} [(\theta_0 - \theta_2)^3 - \theta_1^3] \right].$$

Therefore, ε_t is generally not a martingale difference.

2.2.2. Contemporaneous aggregation and marginalization

In econometrics, for instance, it is often the case that the series of interest results from the aggregation of several (maybe unobserved) other series. Consider the simple case of aggregation of two independent strong MA(1) processes:

$$X_t = X_t^{(1)} + X_t^{(2)}, \quad \text{where } X_t^{(i)} = \eta_t^{(i)} - \theta_i \eta_{t-1}^{(i)}, \quad i = 1, 2.$$

$\theta_1 \neq \theta_2$; the $\eta_t^{(i)}$'s are two independent i.i.d. white noise with third moments not equal to zero. Straightforward calculations show that X_t is also an MA(1) process. However, from computations similar to those of the last example, we show that it is not a strong one. The result can obviously be extended to more general ARMA processes.

A similar problem is that of the marginalization of a strong ARMA vector process. The components are also ARMA processes but they are not strong ones in general.

All these examples have important practical meanings and emphasize the need for estimating weak ARMA representations.

3. Consistency of the least-squares estimator

Let $(X_t)_{t \in \mathbb{Z}}$ be a second-order stationary process such that, for all $t \in \mathbb{Z}$,

$$X_t + \sum_{i=1}^p a_i X_{t-i} = \varepsilon_t + \sum_{i=1}^q b_i \varepsilon_{t-i}, \quad (5)$$

where (ε_t) is a sequence of uncorrelated random variables defined on some probability space (Ω, \mathcal{A}, P) with zero mean and common variance $\sigma^2 > 0$, and where the polynomials $\phi(z) = 1 + a_1 z + \dots + a_p z^p$ and $\psi(z) = 1 + b_1 z + \dots + b_q z^q$ have all their zeros outside the unit disk and have no zero in common. Without loss of generality, assume that a_p and b_q are not both equal to zero (by convention $a_0 = b_0 = 1$). Process (ε_t) can be interpreted as the linear innovation of (X_t) , i.e.

$$\varepsilon_t = X_t - E(X_t / H_X(t-1)),$$

where $H_X(t-1)$ is the Hilbert space generated by $(X_s; s < t)$. In addition, assume that (X_t) is a strictly stationary and ergodic sequence.

The parameter

$$\theta_0 := (a_1, \dots, a_p, b_1, \dots, b_q)',$$

belongs to the parameter space

$$\Theta := \{ \theta = (\theta_1, \dots, \theta_p, \theta_{p+1}, \dots, \theta_{p+q})'; \phi_\theta(z) = 1 + \theta_1 z + \dots + \theta_p z^p \text{ and } \psi_\theta(z) = 1 + \theta_{p+1} z + \dots + \theta_{p+q} z^q \text{ have all their zeros outside the unit disk} \}.$$

For all $\theta \in \Theta$, let $(\varepsilon_t(\theta))$ be the second-order stationary process (see, e.g., Brockwell and Davis, 1991, Chap. 3) for the existence and the uniqueness of such a process) defined as the solution of

$$\varepsilon_t(\theta) = X_t + \sum_{i=1}^p \theta_i X_{t-i} - \sum_{i=1}^q \theta_{p+i} \varepsilon_{t-i}(\theta), \quad \forall t \in \mathbb{Z}. \quad (6)$$

Note that $\varepsilon_t(\theta_0) = \varepsilon_t$ a.s. $\forall t \in \mathbb{Z}$. The assumption on the MA polynomial ψ_θ implies that there exists a sequence of constants $(c_i(\theta))$ such that

$$\sum_{i=1}^{\infty} |c_i(\theta)| < \infty$$

and

$$\varepsilon_t(\theta) = X_t + \sum_{i=1}^{\infty} c_i(\theta) X_{t-i}, \quad \forall t \in \mathbb{Z}. \quad (7)$$

Note that for all $\theta \in \Theta$, $\varepsilon_t(\theta)$ belongs to $L^2(\Omega, \mathcal{A}, P)$, that $(\varepsilon_t(\theta))_{t \in \mathbb{Z}}$ is an ergodic sequence (from (7)), and that $\varepsilon_t(\cdot)$ is a continuous function.

Given a realization of length n , X_1, X_2, \dots, X_n , $\varepsilon_t(\theta)$ can be approximated, for $0 < t \leq n$, by $e_t(\theta)$ defined recursively by

$$e_t(\theta) = X_t + \sum_{i=1}^p \theta_i X_{t-i} - \sum_{i=1}^q \theta_{p+i} e_{t-i}(\theta), \tag{8}$$

where the unknown starting values are set to zero: $e_0(\theta) = e_{-1}(\theta) = \dots = e_{-q+1}(\theta) = X_0 = X_{-1} = \dots = X_{-p+1} = 0$.

Let δ be a strictly positive constant chosen so that the true parameter θ_0 belongs to the compact set

$$\Theta_\delta := \{ \theta \in \mathbb{R}^{p+q}; \text{ the zeros of polynomials } \phi_\theta(z) \text{ and } \psi_\theta(z) \text{ have moduli } \geq 1 + \delta \}.$$

The random variable $\hat{\theta}_n$ is called least-squares estimator if it satisfies, almost surely,

$$Q_n(\hat{\theta}_n) = \min_{\theta \in \Theta_\delta} Q_n(\theta), \tag{9}$$

where

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^n e_t^2(\theta).$$

To prove the consistency and asymptotic normality of the least-squares estimator, it will be convenient to consider the functions

$$O_n(\theta) = \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2(\theta),$$

where $(\varepsilon_t(\theta))$ is given by (7). The first main result obtained in this paper is the following theorem.

Theorem 1. *Let $(X_t)_{t \in \mathbb{Z}}$ be a strictly stationary and ergodic process belonging to L^2 and satisfying (5). Let $(\hat{\theta}_n)$ be a sequence of least-squares estimators defined by (9). Suppose $\theta_0 \in \Theta_\delta$. Then*

$$\hat{\theta}_n \rightarrow \theta_0 \quad \text{a.s. as } n \rightarrow \infty.$$

4. Asymptotic normality

Let $\mathcal{F}_{-\infty}^t$ and \mathcal{F}_{t+k}^∞ be the σ -fields generated by $\{X_u: u \leq t\}$ and $\{X_u: u \geq t+k\}$, respectively. The strong mixing coefficients $(\alpha_X(k))_{k \in \mathbb{N}^*}$ of the stationary process $(X_t)_{t \in \mathbb{Z}}$ are defined by

$$\alpha_X(k) = \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+k}^\infty} |P(A \cap B) - P(A)P(B)|.$$

Pham (1986) has shown that, for a wide class of bilinear processes, the strong mixing coefficient tends to zero exponentially fast. The same property is very easy to

prove for the switching-regime model (3) (X_t is a function of $(\eta_t, \eta_{t-1}, \Delta_t)$ and the strong mixing coefficients of the process $(\eta_t, \eta_{t-1}, \Delta_t)_{t \in \mathbb{Z}}$ tend to zero exponentially fast).

For any $\theta \in \Theta_\delta$, let $\frac{\partial}{\partial \theta} O_n(\theta) = (\frac{\partial}{\partial \theta_1} O_n(\theta), \dots, \frac{\partial}{\partial \theta_{p+q}} O_n(\theta))'$. We consider the following matrices

$$I = \lim_{n \rightarrow \infty} \text{Var} \left(\sqrt{n} \frac{\partial}{\partial \theta} O_n(\theta_0) \right) \quad \text{and} \quad J = \lim_{n \rightarrow \infty} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} O_n(\theta_0) \right] \quad \text{a.s.}$$

the existence of which will be established in the sequel. The second main result of the paper concerns the limiting distribution of $\hat{\theta}_n$.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied. In addition, suppose that $(X_t)_{t \in \mathbb{Z}}$ satisfies $E|X_t|^{4+2\nu} < \infty$ and*

$$\sum_{k=0}^{\infty} [\alpha_X(k)]^{\nu/(2+\nu)} < \infty, \quad (10)$$

for some $\nu > 0$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0)$$

has a limiting centred normal distribution with covariance matrix $J^{-1}IJ^{-1}$.

Remark. The nondegeneracy of the limiting distribution requires that matrix I is positive definite. This is easily shown to be true, for example, in the case where the process (ε_t) is a martingale difference. In particular cases, the property can be checked directly using a characterization of I given in the following section. It requires, however, the computation of expectations of some products of present and past values of X_t .

5. Lemmas and proofs

Proof of Theorem 1. For general vector-valued processes, the proof of consistency has been given in Dunsmuir and Hannan (1976) using spectral analysis. We use a different approach. However, for the sake of brevity we will only sketch the proof. Details are provided in the original version of this paper, which is available on request.

The proof of consistency is mainly based on the ergodic theorem applied to the process $S_m(t) = \inf_{\theta \in V_m(\theta^*) \cap \Theta} \varepsilon_t^2(\theta)$, where $V_m(\theta^*)$ is the open sphere with centre θ^* and radius $1/m$. It also uses a uniform domination result on the coefficients $c_t(\theta)$ on Θ_δ and an asymptotic identifiability result due to the orthogonality properties of the linear innovation $\varepsilon_t(\theta_0)$. Finally, it requires to show that $\varepsilon_t(\theta) - e_t(\theta)$ converges uniformly to zero (a.s.) as t goes to infinity.

The proof of Theorem 2 will be divided into several steps.

Lemma 1. For any $\theta \in \Theta$ and any $m \in \{1, \dots, p+q\}$, there exist absolutely summable sequences $(c_i(\theta))_{i \geq 1}$ and $(c_{i,m}(\theta))_{i \geq 1}$ such that

$$\varepsilon_t(\theta) = X_t + \sum_{i=1}^{\infty} c_i(\theta) X_{t-i} \quad \text{and} \quad \frac{\partial}{\partial \theta_m} \varepsilon_t(\theta) = \sum_{i=1}^{\infty} c_{i,m}(\theta) X_{t-i}. \tag{11}$$

Moreover, there exist $\rho \in [0, 1[$ and $K \in [0, \infty[$ such that, for all $i \geq 1$,

$$\sup_{\theta \in \Theta_\delta} |c_i(\theta)| \leq K \rho^i \quad \text{and} \quad \sup_{\theta \in \Theta_\delta} |c_{i,m}(\theta)| \leq K \rho^i. \tag{12}$$

Proof. To prove the first inequality in (12) we use the fact that the coefficients of Taylor series for $1/\psi_\theta(z)$ (which is absolutely convergent for $|z| < 1 + \delta$) decay at an exponential rate uniformly on Θ_δ . For more details see the original version of this paper, which is available on request.

To prove (11) and the second inequality in (12), first note that they are obvious for $q = 0$. Otherwise, partition θ in $\theta = (\theta^{(1)'}, \theta^{(2)'})'$, where

$$\theta^{(1)'} = (\theta_1, \dots, \theta_p) \quad \text{and} \quad \theta^{(2)'} = (\theta_{p+1}, \dots, \theta_{p+q}).$$

Let $Y_t(\theta^{(1)}) = X_t + \sum_{i=1}^p \theta_i X_{t-i} = \varepsilon_t(\theta) + \sum_{i=1}^q \theta_{i+p} \varepsilon_{t-i}(\theta)$, for all $\theta \in \Theta$. Let $(\tilde{c}_i(\theta^{(2)}))_{i \in \mathbb{N}}$ be the absolutely summable sequence such that $\varepsilon_t(\theta) = \sum_{i=0}^{\infty} \tilde{c}_i(\theta^{(2)}) Y_{t-i}(\theta^{(1)})$. First, consider the case $1 \leq m \leq p$. We have, with probability one

$$\frac{\partial}{\partial \theta_m} \varepsilon_t(\theta) = \sum_{i=0}^{\infty} \tilde{c}_i(\theta^{(2)}) \frac{\partial}{\partial \theta_m} Y_{t-i}(\theta^{(1)}) = \sum_{i=0}^{\infty} \tilde{c}_i(\theta^{(2)}) X_{t-i-m}$$

(the derivation under the sign sum is valid since, with probability one, the series of the derivatives is uniformly absolutely summable in a neighbourhood of θ). Now, consider the case $p+1 \leq m \leq p+q$. Let φ_i , $1 \leq i \leq m$, denote the inverses of the distinct zeros of ψ_θ . Straightforward algebra shows that $\tilde{c}_i(\theta^{(2)})$ can be expressed as a polynomial of the φ_j 's. Therefore, $\tilde{c}_i(\theta^{(2)})$ is differentiable with respect to θ_m . Similarly, it can be shown that $(\partial/\partial \theta_m) \tilde{c}_i(\theta^{(2)})$ is also bounded by a sequence decaying at an exponential rate uniformly on a neighbourhood of θ , $\theta \in \Theta$. Therefore, we have with probability one

$$\frac{\partial}{\partial \theta_m} \varepsilon_t(\theta) = \sum_{i=0}^{\infty} \frac{\partial}{\partial \theta_m} \tilde{c}_i(\theta^{(2)}) Y_{t-i}(\theta^{(1)}),$$

which states (11) with

$$c_{i,m}(\theta) = \frac{\partial}{\partial \theta_m} \tilde{c}_i(\theta^{(2)}) \sum_{j=1}^p \frac{\partial}{\partial \theta_m} \tilde{c}_{i-j}(\theta^{(2)}) \theta_j \quad (\text{by convention } \frac{\partial}{\partial \theta_m} \tilde{c}_i(\theta^{(2)}) = 0 \text{ for } i < 0). \quad \square$$

Lemma 2. For any $\theta \in \Theta$ and any $m \in \{1, \dots, p + q\}$, the random variable

$$\frac{\partial}{\partial \theta_m} O_n(\theta),$$

exists and belongs to L^2 .

Proof. Since, $EX_t^4 < \infty$, $\sum_{i=1}^{\infty} |c_i(\theta)| < \infty$ and $\sum_{i=1}^{\infty} |c_{i,m}(\theta)| < \infty$, by the Cauchy criterion, it can be shown that $E\varepsilon_t^4(\theta) < \infty$ and $E(\partial/\partial\theta_m)\varepsilon_t(\theta)^4 < \infty$. The Cauchy–Schwarz inequality shows that $(\partial/\partial\theta_m)\varepsilon_t^2 = 2\varepsilon_t(\theta)(\partial/\partial\theta_m)\varepsilon_t(\theta)$ belongs to L^2 . The conclusion follows. \square

Lemma 3. Let the assumptions of Theorem 2 be satisfied. For all $\theta \in \Theta_\delta$, the matrix

$$I(\theta) = \lim_{n \rightarrow \infty} \text{Var} \left(\sqrt{n} \frac{\partial}{\partial \theta} O_n(\theta) \right)$$

exists.

Proof. Let $(\partial/\partial\theta)\varepsilon_t(\theta) = ((\partial/\partial\theta_1)\varepsilon_t(\theta), \dots, (\partial/\partial\theta_{p+q})\varepsilon_t(\theta))'$ and $Y_t = 2\varepsilon_t(\theta)(\partial/\partial\theta)\varepsilon_t(\theta)$. We have

$$I_n := \text{Var} \left(\sqrt{n} \frac{\partial}{\partial \theta} O_n(\theta) \right) = \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n \text{Cov}(Y_t, Y_s).$$

For $(l, m) \in \{1, \dots, p + q\}^2$ and $k \in \mathbb{Z}$, let

$$c(k) = \text{cov}(Y_t(l), Y_{t-k}(m)),$$

where $Y_t(l)$ denotes the l th element of Y_t .

Let

$$M(j - i, i' + k - i, j' + k - i) = EX_{t-i}X_{t-j}X_{t-i'-k}X_{t-j'-k} \\ - EX_{t-i}X_{t-j}EX_{t-i'-k}X_{t-j'-k}.$$

First, suppose that $k \geq 0$. From the dominated convergence theorem we have

$$|c(k)| = 4 \left| \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{i'=0}^{\infty} \sum_{j'=0}^{\infty} \text{cov}(c_i(\theta)X_{t-i}c_{j,l}(\theta)X_{t-j}, c_{i'}(\theta)X_{t-i'-k}c_{j',m}(\theta)X_{t-j'-k}) \right| \\ = 4 \left| \sum_{i,j,i',j'} c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j - i, i' + k - i, j' + k - i) \right| \\ \leq 4(g_1 + g_2 + g_3 + g_4 + g_5),$$

where

$$\begin{aligned}
 g_1 &= \sum_{i>k/2} \sum_{j,i',j'=0}^{\infty} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j-i,i'+k-i,j'+k-i)|, \\
 g_2 &= \sum_{i'>k/2} \sum_{i,j,j'=0}^{\infty} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j-i,i'+k-i,j'+k-i)|, \\
 g_3 &= \sum_{j>k/2} \sum_{i,i',j'=0}^{\infty} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j-i,i'+k-i,j'+k-i)|, \\
 g_4 &= \sum_{j'>k/2} \sum_{i,i',j=0}^{\infty} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j-i,i'+k-i,j'+k-i)|, \\
 g_5 &= \sum_{i\leq k/2, j\leq k/2, i'\leq k/2, j'\leq k/2} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M(j-i,i'+k-i,j'+k-i)|.
 \end{aligned}$$

The Cauchy–Schwarz inequality implies that

$$\begin{aligned}
 |M(j-i,i'+k-i,j'+k-i)| &\leq \sqrt{E(X_{t-i}X_{t-j})^2 E(X_{t-i'-k}X_{t-j'-k})^2} \\
 &\leq EX_t^4 = M < \infty.
 \end{aligned}$$

Therefore, using Lemma 1, we have

$$\begin{aligned}
 g_1 &\leq \sum_{i,j,i',j' \text{ and } i>k/2} |c_i(\theta)c_{j,l}(\theta)c_{i'}(\theta)c_{j',m}(\theta)M| \\
 &\leq M \sum_{i>k/2} |c_i(\theta)| \sum_j |c_{j,l}(\theta)| \sum_{i'} |c_{i'}(\theta)| \sum_{j'} |c_{j',m}(\theta)| \\
 &\leq M_1 \sum_{i>k/2} |c_i(\theta)| \\
 &\leq M_2 \rho^{k/2},
 \end{aligned}$$

for some positive constants M_1 and M_2 . The same inequality holds for g_2, g_3 and g_4 . Note that $E|X_t X_s|^{2+\nu} < \infty$ for some $\nu > 0$. The Davydov inequality (Davydov, 1968) implies that, for $0 \leq i, j, i', j' \leq k/2$, there exist positive constants C and C_1 such that

$$\begin{aligned}
 &M(j-i,i'+k-i,j'+k-i) \\
 &\leq C \|X_{t-i}X_{t-j}\|_{2+\nu} \|X_{t-i'-k}X_{t-j'-k}\|_{2+\nu} \\
 &\quad \times (\alpha_X(\min\{k+i'-j, k+j'-j, k+i'-i, k+j'-i\}))^{\nu/(2+\nu)} \\
 &\leq C_1 \left(\alpha_X \left(\left[\frac{k}{2} \right] \right) \right)^{\nu/(2+\nu)}.
 \end{aligned}$$

Therefore, there exists a positive constant M_3 such that

$$\begin{aligned}
 g_5 &\leq C_1 \left(\alpha_X \left(\left[\frac{k}{2} \right] \right) \right)^{\nu/(2+\nu)} \left(\sum_i |c_i(\theta)| \right)^2 \sum_j |c_{j,l}(\theta)| \sum_j |c_{j,m}(\theta)| \\
 &\leq M_3 \left(\alpha_X \left(\left[\frac{k}{2} \right] \right) \right)^{\nu/(2+\nu)}.
 \end{aligned}$$

Thus, for $k \geq 0$, we have

$$|c(k)| \leq 4M_2 \rho^{|k|/2} + M_3 \left(\alpha_X \left(\left[\frac{|k|}{2} \right] \right) \right)^{v/(2+v)}.$$

A similar inequality holds for $k \leq 0$. Therefore (10) implies that

$$\sum_{k=-\infty}^{\infty} |c(k)| < \infty.$$

Then the dominated convergence theorem gives

$$\begin{aligned} I_n(l, m) &= \frac{1}{n} \sum_{-n < k < n} (n - |k|)c(k) \\ &\rightarrow \sum_{k=-\infty}^{\infty} c(k), \end{aligned}$$

as $n \rightarrow \infty$. \square

Lemma 4. *Under the assumptions of Theorem 2 the random vector $\sqrt{n}(\partial/\partial\theta)Q_n(\theta_0)$ has a limiting normal distribution with mean 0 and covariance matrix $I(\theta_0)$.*

Proof. It is easy to show that $\sqrt{n}(\partial/\partial\theta)(Q_n(\theta_0) - O_n(\theta_0))$ converges in probability to zero. Therefore, $\sqrt{n}(\partial/\partial\theta)Q_n(\theta_0)$ and $\sqrt{n}(\partial/\partial\theta)O_n(\theta_0)$ have the same asymptotic distribution. Note that $E_{\theta_0}(\partial/\partial\theta)\varepsilon_t^2(\theta_0) = 2E_{\theta_0}\varepsilon_t(\partial/\partial\theta)\varepsilon_t(\theta_0) = 0$, since $(\partial/\partial\theta)\varepsilon_t(\theta)$ belongs to $H_X(t - 1)$. Therefore, $\sqrt{n}(\partial/\partial\theta)O_n(\theta_0)$ is centred. Maintain the notations of the proof of Lemma 3. For any positive integer r , we have

$$\sqrt{n} \frac{\partial}{\partial\theta} O_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{Y}_t = \frac{1}{\sqrt{n}} \sum_{t=1}^n (\mathbf{Y}_{t,r} - E_{\theta_0} \mathbf{Y}_{t,r}) + \frac{1}{\sqrt{n}} \sum_{t=1}^n (\mathbf{Z}_{t,r} - E_{\theta_0} \mathbf{Z}_{t,r})$$

where

$$\mathbf{Z}_{t,r} = 2\mathbf{U}_{t,r} + 2\mathbf{V}_{t,r},$$

$$\mathbf{U}_{t,r} = \sum_{i=r+1}^{\infty} c_i(\theta_0) X_{t-i} \sum_{j=1}^{\infty} X_{t-j} (c_{j,1}(\theta_0), \dots, c_{j,p+q}(\theta_0))',$$

$$\mathbf{V}_{t,r} = \sum_{i=0}^r c_i(\theta_0) X_{t-i} \sum_{j=r+1}^{\infty} X_{t-j} (c_{j,1}(\theta_0), \dots, c_{j,p+q}(\theta_0))',$$

and

$$\mathbf{Y}_{t,r} = 2 \sum_{i=0}^r c_i(\theta_0) X_{t-i} \sum_{j=1}^r X_{t-j} (c_{j,1}(\theta_0), \dots, c_{j,p+q}(\theta_0))'.$$

First note that $\mathbf{Y}_{t,r}$ is a function of a finite number of values of the process (X_t) . Therefore, the stationary process $(\mathbf{Y}_{t,r})_{t \in \mathbb{Z}}$ satisfies a mixing property of the form (10). The central limit theorem for strongly mixing processes (Ibragimov, 1962) implies that $(1/\sqrt{n}) \sum_{t=1}^n (\mathbf{Y}_{t,r} - E_{\theta_0} \mathbf{Y}_{t,r})$ has a limiting $\mathcal{N}(0, \tilde{I}_r)$ distribution. Moreover, standard

calculations show that $\tilde{I}_r \rightarrow I$ as $r \rightarrow \infty$. Now we will show that $E(\frac{1}{\sqrt{n}} \sum_{t=1}^n (\mathbf{Z}_{t,r} - E_{\theta_0} \mathbf{Z}_{t,r})) (\frac{1}{\sqrt{n}} \sum_{t=1}^n (\mathbf{Z}_{t,r} - E_{\theta_0} \mathbf{Z}_{t,r}))'$ converges to 0 uniformly in n as $r \rightarrow \infty$. The conclusion will follow from a straightforward adaptation of a result given by Anderson (Corollary 7.7.1, 1971, p. 426).

For $m \in \{1, \dots, p + q\}$ we have

$$\begin{aligned} \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n U_{t,r}(m) \right) &= \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n \text{cov}(U_{t,r}(m), U_{s,r}(m)) \\ &= \frac{1}{n} \sum_{-n < h < n} (n - |h|) c_r(h) \\ &\leq \sum_{h=-\infty}^{\infty} |c_r(h)|, \end{aligned}$$

where

$$\begin{aligned} c_r(h) &= \text{cov}(U_{t,r}(m), U_{t+h,r}(m)) \\ &= \sum_{i=r+1}^{\infty} \sum_{j=1}^{\infty} \sum_{i'=r+1}^{\infty} \sum_{j'=1}^{\infty} c_i(\theta) c_{i'}(\theta) c_{j,m}(\theta) c_{j',m}(\theta) \text{cov}(X_{t-i} X_{t-j}, X_{t+h-i'} X_{t+h-j'}). \end{aligned}$$

First, suppose that $h \geq 0$. Using the Cauchy–Schwarz inequality and the Davydov inequality, we show that there exists a positive constant C such that

$$|c_r(h)| \leq M \sum_{i=r+1}^{\infty} |c_i(\theta)| \sum_{j=1}^{\infty} |c_{j,m}(\theta)| \sum_{i'=r+1}^{\infty} |c_{i'}(\theta)| \sum_{j'=1}^{\infty} |c_{j',m}(\theta)| \leq C \rho^r$$

and, for $r < [h]/2$,

$$\begin{aligned} |c_r(h)| &\leq \sum_{r < i' < [h/2]} \sum_{0 < j' < [h/2]} \sum_i \sum_j |c_i(\theta) c_{i'}(\theta) c_{j,m}(\theta) c_{j',m}(\theta)| \\ &\quad \times |\text{cov}(X_{t-i} X_{t-j}, X_{t+h-i'} X_{t+h-j'})| \\ &\quad + \sum_{i' \geq [h/2]} \sum_{i,j,j'} |c_i(\theta) c_{i'}(\theta) c_{j,m}(\theta) c_{j',m}(\theta)| |\text{cov}(X_{t-i} X_{t-j}, X_{t+h-i'} X_{t+h-j'})| \\ &\quad + \sum_{j' \geq [h/2]} \sum_{i,j,i'} |c_i(\theta) c_{i'}(\theta) c_{j,m}(\theta) c_{j',m}(\theta)| |\text{cov}(X_{t-i} X_{t-j}, X_{t+h-i'} X_{t+h-j'})| \\ &\leq C \rho^r \left(\alpha_X \left(\left[\frac{|h|}{2} \right] \right) \right)^{v/(2+v)} + C \rho^r \rho^{|h|/2}. \end{aligned}$$

The same inequality holds for $h < 0$. Therefore, there exists a constant C_1 such that

$$\begin{aligned} \sum_{h=-\infty}^{\infty} |c_r(h)| &= \sum_{|h| \leq 2r+1} |c_r(h)| + \sum_{|h| \geq 2(r+1)} |c_r(h)| \leq C_1 r \rho^r + C_1 \rho^r \\ &\quad + C_1 \rho^r \sum_k (\alpha_X(k))^{v/(2+v)} \rightarrow 0 \end{aligned}$$

as $r \rightarrow \infty$. Thus,

$$\sup_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n U_{t,r}(m) \right) \rightarrow 0$$

as $r \rightarrow \infty$. Similarly, it can be shown that

$$\sup_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n V_{t,r}(m) \right) \rightarrow 0$$

as $r \rightarrow \infty$. Then

$$\begin{aligned} \sup_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_{t,r}(m) \right)^2 &\leq 8 \sup_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n U_{t,r}(m) \right)^2 \\ &+ 8 \sup_n \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n V_{t,r}(m) \right)^2 \rightarrow 0 \end{aligned}$$

as $r \rightarrow \infty$ and the proof is complete. \square

Lemma 5. *Almost surely the matrix J exists and is strictly positive definite.*

Proof. It is easy to show that $|(\partial^2/\partial\theta_i\partial\theta_j)\varepsilon_t(\theta) - (\partial^2/\partial\theta_i\partial\theta_j)\varepsilon_t(\theta_0)| \rightarrow 0$ almost surely, as $t \rightarrow \infty$. Therefore, $(\partial^2/\partial\theta_i\partial\theta_j)O_n(\theta_0)$ and $(\partial^2/\partial\theta_i\partial\theta_j)Q_n(\theta_0)$ have almost surely the same limit (when existing). As in the proof of Lemma 1, it can be shown that

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\varepsilon_t(\theta) = \sum_{l=1}^{\infty} c_{l,i,j}(\theta)X_{t-l},$$

where $\sum_{l=1}^{\infty} |c_{l,i,j}(\theta)| < \infty$. Therefore, $(\partial^2/\partial\theta_i\partial\theta_j)\varepsilon_t(\theta_0)$ belongs to L^2 . Then we have

$$\begin{aligned} \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}Q_n(\theta_0) \right] &= \frac{2}{n} \left[\sum_{t=1}^n \varepsilon_t \frac{\partial^2}{\partial\theta_i\partial\theta_j}\varepsilon_t(\theta_0) + \frac{\partial}{\partial\theta_i}\varepsilon_t(\theta_0) \frac{\partial}{\partial\theta_j}\varepsilon_t(\theta_0) \right] \\ &\rightarrow 2 \left[E\varepsilon_t \frac{\partial^2}{\partial\theta_i\partial\theta_j}\varepsilon_t(\theta_0) + E \frac{\partial}{\partial\theta_i}\varepsilon_t(\theta_0) \frac{\partial}{\partial\theta_j}\varepsilon_t(\theta_0) \right]. \end{aligned}$$

Since $(\partial^2/\partial\theta_i\partial\theta_j)\varepsilon_t(\theta)$ belongs to $H_X(t-1)$, we have $E\varepsilon_t(\partial^2/\partial\theta_i\partial\theta_j)\varepsilon_t(\theta_0) = 0$. Therefore, J is the covariance matrix of $\sqrt{2}(\partial/\partial\theta)\varepsilon_t(\theta_0)$. If it is not strictly positive definite then there exist real constants β_m , not all equal to zero, such that $\sum_{m=1}^{p+q} \beta_m \frac{\partial}{\partial\theta_m}\varepsilon_t(\theta_0) = \sum_{l=1}^{\infty} (\sum_{m=1}^{p+q} \beta_m c_{l,m})X_{t-l} = 0$. This is in contradiction with the assumption that the variance σ^2 of the linear innovations is not equal to zero. \square

Proof of Theorem 2. Using a standard technique of Taylor expansion around θ_0 , we obtain

$$0 = \sqrt{n} \frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n) = \sqrt{n} \frac{\partial}{\partial\theta}Q_n(\theta_0) + \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}Q_n(\theta_{n,i,j}^*) \right] \sqrt{n}(\hat{\theta}_n - \theta_0),$$

where the $\theta_{n,i,j}^*$'s are between $\hat{\theta}_n$ and θ_0 . Doing again a Taylor expansion we obtain

$$\left| \frac{\partial^2}{\partial\theta_i\partial\theta_j} Q_n(\theta_{n,i,j}^*) - \frac{\partial^2}{\partial\theta_i\partial\theta_j} Q_n(\theta_0) \right| \leq \sup_{\theta \in \Theta_b} \left\| \frac{\partial}{\partial\theta} \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} Q_n(\theta) \right) \right\| \|\theta_{n,i,j}^* - \theta_0\| \rightarrow 0$$

almost surely as $n \rightarrow \infty$. From Lemmas 5 and 4, we obtain that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ has a limiting normal distribution with mean 0 and covariance matrix $J^{-1}IJ^{-1}$. \square

6. Examples and numerical illustration

We first discuss a number of examples for which the asymptotic variance of the estimator can be given in closed form.

6.1. Covariance matrix calculations

Example 1 (noncausal AR(1)). Consider the process: $X_t = -\sum_{j=1}^{\infty} \phi^{-j} \eta_{t+j}$, where $|\phi| > 1$ and (η_t) is a non-Gaussian i.i.d. white noise. Then (X_t) admits the following noncausal AR(1) representation: $X_t = \phi X_{t-1} + \eta_t$. Moreover, it is well known that the following AR(1) representation holds: $X_t = (1/\phi)X_{t-1} + \varepsilon_t$, where (ε_t) is a weak white noise. If, for instance, $E(\eta_t^3) \neq 0$, then it is easily checked that the (ε_t) process is not a martingale difference. Now accounting for the weak AR(1) representation of (X_t) , the least-squares estimator of parameter $1/\phi$ can be considered. According to Theorem 2, some moment calculations show that the asymptotic variance $J^{-1}IJ^{-1}$ is equal to $(\phi^2 - 1)/\phi^2$. The same asymptotic variance is obtained by assuming (ε_t) is i.i.d. This is not surprising since the least-squares estimator of $1/\phi$ has the same asymptotic distribution as the first-order empirical autocorrelation of (X_t) . Further, the Bartlett formula applies regardless of the hypothesis on the white noise (weak or strong). Therefore, there is no loss of asymptotic efficiency due to the dependence structure of (ε_t) . The next example shows that it is not always the case.

Example 2 (switching-regime model). In Section 2, we exhibited an MA(1) representation for a switching-regime Markov model. For simplicity, assume that $\eta_t \sim \mathcal{N}(0,1)$, $b = -a \in]-1, 0]$ and $p := P(\Delta_t = 1/\Delta_{t-1} = 0) = P(\Delta_t = 0/\Delta_{t-1} = 1) \in]0, 1[$. In this case (X_t) is a weak white noise. Therefore, (X_t) satisfies the following weak MA(1) representation: $X_t = \varepsilon_t + c\varepsilon_{t-1}$, where $c = 0$ and $(\varepsilon_t) = (X_t)$ is a weak white noise with variance $E\varepsilon_t^2 = 1 + a^2$. It follows from Theorem 2 and standard computations involving moments of X_t , that the asymptotic variance is equal to

$$J^{-1}IJ^{-1} = 1 + \frac{2a^2(1 - 2p)}{(1 + a^2)^2} + \frac{a^2}{p(1 + a^2)^2}. \tag{13}$$

It should be noted that the asymptotic covariance matrix would equal 1 if the MA(1) representation was a strong one. Therefore, a huge discrepancy between the asymptotic

precisions of the strong and weak representations can hold (for p close to zero). A more astonishing output of this computation is that the weak representation can be more easily estimated (asymptotically) than the strong one: choosing p close to 1 in (13) shows that the asymptotic variance can be less than 1.

Example 3 (bilinear). Consider the following bilinear equation:

$$\forall t \in \mathbb{Z}, \quad X_t = \eta_t + b\eta_{t-1}X_{t-2}, \tag{14}$$

where (η_t) is an i.i.d. $\mathcal{N}(0, 1)$ sequence, and where b is a real constant such that $3b^4 < 1$. Under the previous assumptions on the coefficient b and the sequence (η_t) , it can be shown that (14) admits a unique stationary nonanticipative solution $X = (X_t)_{t \in \mathbb{Z}}$ (see, e.g., Guégan, 1981). Moreover, X admits fourth-order moments and it is a weak white noise.

Now, suppose that the statistician is mistaking it for an AR(1) process. Then he will estimate the following representation:

$$X_t = \varepsilon_t + aX_{t-1}, \quad \forall t \in \mathbb{Z}, \tag{15}$$

where $a = 0$ is the true value. The least-squares estimator is given by

$$\hat{a} = \frac{\sum_{t=2}^n X_t X_{t-1}}{\sum_{t=2}^n X_{t-1}^2}. \tag{16}$$

Since the nonanticipative stationary solution of (14) is ergodic, we are able to check directly in this case the consistency stated in Theorem 1:

$$\hat{a} \rightarrow \frac{EX_t X_{t-1}}{EX_t^2} = \frac{\gamma(1)}{\gamma(0)} = 0 = a.$$

Straightforward calculations of the matrices involved in Theorem 2 show that the asymptotic variance of $\sqrt{n}\hat{a}$ is equal to $1 + 4b^2 - 2b^4$. As a comparison, note that the value of $\lim_{n \rightarrow \infty} \text{var}(\sqrt{n}\hat{a})$ is equal to 1 for a strong white noise. Once again, e.g., for $|b| \geq 0.55$, the asymptotic variance of $\sqrt{n}\hat{a}$ is more than twice as big as the corresponding asymptotic variance in the case of a strong white noise. Therefore, it is seen that, using standard confidence intervals, the statistician is likely to mistake the hypothesis that $a \neq 0$ as being true.

6.2. Covariance matrix estimation

It is clear from the previous examples that the asymptotic covariance matrix $J^{-1}JJ^{-1}$ of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be very different from the one obtained for a strong ARMA model. We now consider estimation of this matrix.

As in some of the previous examples the least-squares estimator can sometimes be approximated by a differentiable function of a finite number of empirical autocovariances:

$$\hat{\theta}_n = g(\hat{\gamma}(0), \hat{\gamma}(1), \dots, \hat{\gamma}(m)) + o_p(1),$$

where g is continuously differentiable in a neighbourhood of $(\gamma(0), \gamma(1), \dots, \gamma(m))$ and $o_p(1)$ denotes a sequence of random vectors which converges in probability to zero. Under the assumptions of Theorem 2 the vector $\sqrt{n}((\hat{\gamma}(0), \hat{\gamma}(1), \dots, \hat{\gamma}(m))' - (\gamma(0), \gamma(1), \dots, \gamma(m))')$ is asymptotically normally distributed with zero mean and covariance matrix Σ defined by

$$\Sigma(i + 1, j + 1) = \lim_{n \rightarrow \infty} n \operatorname{cov}(\hat{\gamma}(i), \hat{\gamma}(j)) = \sum_{k=-\infty}^{\infty} \sigma_{i,j}(k) \quad \text{for } (i, j) \in \{0, 1, \dots, m\}^2,$$

where $\sigma_{i,j}(k) = \operatorname{cov}(X_t X_{t-i}, X_{t-k} X_{t-k-j})$. Then the asymptotic covariance matrix of $\sqrt{n}\hat{\theta}_n$ can be expressed as a function of g and Σ , namely $D\Sigma D'$, where D is the matrix of the partial derivatives of g evaluated at $(\gamma(0), \gamma(1), \dots, \gamma(m))$. Therefore, a strongly consistent estimator of D is obtained by plugging the sample autocovariances into the expression of D . Finally, a consistent estimator of $J^{-1}IJ^{-1}$ is obtained from a consistent estimator $\hat{\Sigma}$ of Σ .

For example, consider the AR(1) case. We have

$$\begin{aligned} \hat{\theta}_n = \hat{a} &= \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} + o_p(1), \quad D = \left(-\frac{\gamma(1)}{(\gamma(0))^2}, \frac{1}{\gamma(0)} \right), \\ \hat{D} &= \left(-\frac{\hat{\gamma}(1)}{(\hat{\gamma}(0))^2}, \frac{1}{\hat{\gamma}(0)} \right), \quad J^{-1}\widehat{IJ}^{-1} = \hat{\sigma}_a^2 = \hat{D}\hat{\Sigma}\hat{D}'. \end{aligned}$$

It remains to define a consistent estimator of Σ . For linear processes, Σ can be derived from the Bartlett formula. In this case, Robinson (1977), Mélard et al. (1991) have proposed methods to estimate Σ in a consistent way. For nonlinear processes, the fourth-order moments $\sigma_{i,j}(k)$ involved in the expression of Σ can be approximated by

$$\hat{\sigma}_{i,j}(k) = \frac{1}{n} \sum_{t=1}^{n-i-k} (X_t X_{t-i} - \hat{\gamma}(i))(X_{t-k} X_{t-k-j} - \hat{\gamma}(j)) \quad \text{for } i \leq j \text{ and } 0 \leq k < n - j.$$

The $\hat{\sigma}_{i,j}(k)$'s are defined similarly in the cases $i > j$ and $-n + i < k < 0$. Consider the smoothed empirical estimator of Σ defined by

$$\hat{\Sigma}(i + 1, j + 1) = \sum_{|k| \leq a/b_n} \left(1 - \frac{|k|}{n} \right) \omega(kb_n) \hat{\sigma}_{i,j}(k),$$

where $(b_n)_{n \in \mathbb{N}^*}$ is a sequence of real numbers such that $b_n \rightarrow 0$ and $\sqrt{nb_n} \rightarrow \infty$ as $n \rightarrow \infty$, the function $\omega: \mathbb{R} \rightarrow \mathbb{R}$ is bounded, nonnegative definite with compact support $[-a, a]$, continuous at the origin with $\omega(0) = 1$ and satisfies

$$b_n \sum_{-n < i < n} |\omega(ib_n)| = O(1).$$

For the applications given in this paper we have chosen $b_n = n^{-1/4}$ and $\omega(x) = \max\{0, 1 - |x|\}$. Berline and Francq (1994) have shown that, under the assumption of Theorem 2 and the additional assumption that $EX^{8+4\nu} < \infty$, the estimator $\hat{\Sigma}$ converges in mean square to Σ and is a nonnegative definite matrix. Therefore, we obtain an estimator of $J^{-1}IJ^{-1}$ which converges in probability.

Next, we briefly discuss estimation of matrices I and J in the general case. In the proof of Theorem 2, the following expressions were obtained:

$$I = 4 \sum_{k=-\infty}^{\infty} \text{cov} \left(\varepsilon_t(\theta_0) \frac{\partial}{\partial \theta} \varepsilon_t(\theta_0), \varepsilon_{t-k}(\theta_0) \frac{\partial}{\partial \theta} \varepsilon_{t-k}(\theta_0) \right)$$

and

$$J = 2E \left(\frac{\partial}{\partial \theta} \varepsilon_t(\theta_0) \frac{\partial}{\partial \theta'} \varepsilon_t(\theta_0) \right).$$

Now, given a particular invertible ARMA model which we aim to estimate, the expansions of ε_t and its derivatives as linear combinations of the present and past values of X_t can be used. For example, in the MA(1) case, straightforward computations provide

$$I = 4 \sum_{k=-\infty}^{\infty} \sum_{i_1, i_3=0}^{\infty} \sum_{i_2, i_4=1}^{\infty} (-\theta_0)^{i_1+i_2+i_3+i_4-2} i_2 i_4 E X_{t-i_1} X_{t-i_2} X_{t-i_3} X_{t-i_4},$$

while matrix J can be easily expressed in terms of the autocovariance function of X_t . Then an estimator of matrices I and J can be obtained by plugging suitably weighted (as in the previous example for $\hat{\Sigma}$) fourth-order empirical moments of the observed process into the expansions. The consistency of such estimators, however, is an issue for future researches.

6.3. Numerical illustrations

We consider two numerical illustrations of the estimation procedure. The first one is based on simulated data, while the second one uses real data.

We carried out simulations for all three examples presented in 6.1, which confirm the theoretical results. For sake of brevity, we only present the bilinear case. We generated 1000 independent trajectories of size 500 of the bilinear process (14), with $b=0.5$, using the NAG Fortran workstation library. For each trajectory, an AR(1) is fitted using the NAG routine G13AFF. This routine uses a least-squares procedure incorporating backforecasting. As expected, the estimates hence obtained are very close to those given by (16) (i.e., without backforecasting). Over the 1000 simulations, \hat{a} varies from -0.21 to 0.23 . The observed mean of the 1000 estimates of \hat{a} is -0.002 (while the true value of the parameter is 0). The observed standard deviation of the 1000 estimates of \hat{a} is 0.06 (the asymptotic theoretical distribution gives $\sqrt{1.875/500} = 0.0612$). The NAG routine also provides an estimate of the standard deviation $\sigma_a := \sqrt{\text{Var}_{as}(\sqrt{n}\hat{a})}$. However, the consistency of the NAG routine estimator requires assumptions (see e.g. Brockwell and Davis, 1991, p. 259) which are not satisfied by the simulated process (14). Fig. 1 shows that the estimates of σ_a are generally very close to 1 (which is the value of σ_a for a strong white noise), whereas the true value is $\sigma_a^2 = 1.875$. This can lead to a serious misspecification of the time-series model. Let us test the null hypothesis H_0 : “ X is a weak white noise”. We reject H_0 when the absolute value of \hat{a} is greater than 1.96 times the estimated standard deviation of \hat{a} . If the standard

method	minimal observed value	observed mean	maximal observed value	observed standard deviation
standard procedure for strong ARMA (NAG routine)	0.95	1.00	1.04	0.007
procedure adapted for ARMA representations of non linear processes	0.69	1.58	8.77	0.62

Fig. 1. Comparison between estimates of $\sigma_a^2 := \text{Var}_{as}(\sqrt{n}\hat{a})$ obtained with the standard method and those obtained with a method adapted to mixing non linear processes (the true value of σ_a^2 is 1.875).

deviation is accurately estimated then the error of the first kind is approximately 5%. Using the standard method given by the NAG routine, we get a rejection frequency of H_0 of 14,1% for 1000 trajectories, which is too high.

Our application on real data concerns the well-known data set of Wolfer sunspot numbers ranging from 1770 to 1869. Various parametric time series have been proposed to fit these data; see Tong (1990) for an extensive review. For the ARMA modelling, the AR(2) model is generally selected (see, e.g., Box and Jenkins, 1970). The NAG routine G13AFF gives

$$Y_t - 1.42Y_{t-1} + 0.73Y_{t-2} = \varepsilon_t, \quad (17)$$

where $Y_t = X_t - 47.011$, $t = 1, \dots, 100$ denote the mean-corrected series. The standard deviations corresponding to the two parameters are both equal to 0.07. They are calculated under the assumption that the linear innovations ε_t are independent. However, many studies have shown that this assumption can be seriously questioned. The estimation procedure developed in this paper provides the following fitted model:

$$Y_t - 1.4Y_{t-1} + 0.7Y_{t-2} = \varepsilon_t, \quad (18)$$

where the estimated standard deviations are now equal to 0.2. Therefore, although in this case the AR(2) model is not invalidated, it seems that the standard analysis is too optimistic about the precision of the estimator. Finally, the fitted linear model could serve as a benchmark for selecting a particular nonlinear model compatible with the AR(2) representation.

7. Conclusions

In this paper we gave theoretical results aimed to justify the common practice of fitting ARMA models to possibly nonlinear data sets. Replacing the usual implicit strong assumptions on the noise process by ergodicity and mixing modifies the asymptotic results. Although the estimator remains consistent and asymptotically normal, the asymptotic variance is likely to be affected by the dependence structure imposed on the (linear) innovation process. The empirical study proposed in this paper shows that the modification can be dramatic. Standard identification routines based on strong

hypothesis on the innovation of ARMA models can therefore lead to serious misspecification when these assumptions do not hold: they will result in inappropriate parameter standard errors, and these will typically be too small.

Acknowledgements

We would like to thank an editor and two anonymous referees of this journal for helpful comments and suggestions.

References

- Amemiya, T., Wu, R.Y., 1972. The effect of Aggregation on prediction in the autoregressive model. *J. Amer. Statist. Assoc.* 67, 628–632.
- Anderson, T.W., 1971. *The Statistical Analysis of Time Series*. Wiley, New York.
- Bartlett, M.S., 1955. *An Introduction to Stochastic Processes*. Cambridge University Press, Cambridge.
- Berlinet, A., Francq, C., 1994. Estimating the covariance between two samples autocovariances. *Trans. 12th Prague Conf. Academy of Science of the Czech Republic, Prague*, 35–38.
- Box, G.E.P., Jenkins, G.M., 1970. *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.
- Brockwell, P.J., Davis, R.A., 1988. Simple Consistent Estimation of the Coefficients of a Linear Filter. *Stochastic Process. Appl.* 22, 47–59, Springer, Berlin.
- Brockwell, P.J., Davis, R.A., 1991. *Time Series: Theory and Methods*. Springer, Berlin.
- Davydov, Y.A., 1968. Convergence of distributions generated by stationary stochastic processes. *Theor. Probab. Appl.* 13, 691–696.
- Dunsmuir, W., Hannan, E.J., 1976. Vector linear time series models. *Adv. Appl. Probab.* 8, 339–364.
- Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.
- Francq, C., Zakoïan, J.M., 1995. Multivariate ARMA models with generalized autoregressive linear innovation. Document de travail du CREST no. 9529, INSEE.
- Gouriéroux, C., Monfort, A., 1992. Qualitative threshold ARCH models. *J. Econometrics* 52, 159–199.
- Granger, C.W.J., Andersen, A., 1978. *An Introduction to Bilinear Time Series Models*. Vanderhoeck and Reprecht, Gottingen.
- Guégan, D., 1981. Etude d'un modèle non linéaire, le modèle superdiagonal d'ordre un. *CRAS Série I*, vol. 293, pp. 95–98.
- Guégan, D., 1988. Modèles bilinéaires et polynomiaux de séries chronologiques: étude probabiliste et statistique. Thèse d'état de l'université Grenoble 1.
- Hamilton, J.D., 1994. *Time Series Analysis*. Princeton University Press.
- Hannan, E.J., 1975. The estimation of ARMA models. *Ann. Statist.* 3, 975–981.
- Ibragimov, I.A., 1962. Some limit theorems for stationary processes. *Theory Probab. Appl.* 7, 349–382.
- Klimko, L.A., Nelson, P.I., 1978. On conditional least squares estimation for stochastic processes. *Ann. Statist.* 6, 629–642.
- Lai, T.L., Wei, C.Z., 1983. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *J. Multivariate Anal.* 13, 1–23.
- Li, W.K., McLeod, A.I., 1988. ARMA modelling with non-Gaussian innovations. *J. Time Ser. Anal.* 9, 155–168.
- Mélard, G., Paesmans, M., Roy, R., 1991. Consistent estimation of the asymptotic covariance structure of multivariate serial correlations. *J. Time Ser. Anal.* 12, 351–361.
- Nijman, T.E., Palm, F.C., 1990. Parameter identification in ARMA-processes in the presence of regular but incomplete sampling. *J. Time Ser. Anal.* 11, 239–248.
- Palm, F.C., Nijman, T.E., 1986. Missing observations in the dynamic regression model. *Econometrica* 52, 1415–1435.

- Pham, D.T., 1984. The estimation of parameters for autoregressive moving average models. *J. Time Ser. Anal.* 5, 53–68.
- Pham, D.T., 1985. Bilinear Markovian representation and bilinear models. *Stochastic Process. Appl.* 20, 295–306.
- Pham, D.T., 1986. The mixing property of bilinear and generalized random coefficient autoregressive models. *Stochastic Process. Appl.* 23, 291–300.
- Priestley, M.B., 1988. *Non Linear and Non Stationary Time Series*. Academic Press, New York.
- Robinson, P.M., 1977. Estimating variances and covariances of sample autocorrelations and autocovariances. *Aust. J. Statist.* 19, 236–240.
- Subba Rao, T., 1978. On the Estimation of Parameters of Bilinear Time Series Models. Tech. Rep. °79, Department of Mathematics, University of Manchester Institute of Science and Technology.
- Tiao, G.C., Tsay, R.S., 1983. Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Ann. Statist.* 11, 856–871.
- Tong, H., 1990. *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press Oxford.
- Tong, H., Lim, K.S., 1980. Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. Ser. B*, 42, 245–292.
- Whittle, P., 1963. *Prediction and Regulation*. The English Universities Press, London.
- Yohai, V.J., Maronna, R.A., 1977. Asymptotic behavior of least-squares estimates for autoregressive processes with infinite variances. *Ann. Statist.* 5, 554–560.