Chapter 12

RECENT RESULTS FOR LINEAR TIME SERIES MODELS WITH NON INDEPENDENT INNOVATIONS

Christian Francq
Jean-Michel Zakoïan

Abstract In this paper, we provide a review of some recent results for ARMA models with uncorrelated but non independent errors. The standard so-called Box-Jenkins methodology rests on the errors independence. When the errors are suspected to be non independent, which is often the case in real situations, this methodology needs to be adapted. We study in detail the main steps of this methodology in the above-mentioned framework.

1. Introduction

Among the great diversity of stochastic models for time series, it is customary to make a sharp distinction between linear and non-linear models. It is often argued that linear ARMA models do not allow to deal with data which exhibit nonlinearity features. This is true if, as is usually done, strong assumptions are made on the noise governing the ARMA equation. There are numerous data sets from the fields of biology, economics and finance, which appear to be hardly compatible with these so-called strong ARMA models (see Tong (1982) for examples). When these assumptions are weakened, it can be shown that many non linear processes admit a finite order ARMA representation. A striking feature of the ARMA class is that it is sufficiently flexible and rich enough to accommodate a variety of non-linear models. In fact, these two classes of processes, linear and non-linear, are not incompatible and can even be complementary.

A key question, of course, is how to fit linear ARMA models to data which exhibit non-linear behaviours. The validity of the different steps of the traditional methodology of Box and Jenkins, identification, estimation and validation, crucially depends on the noise properties. If the noise is not independent, most of the tools (such as parameters confidence intervals, autocorrelation significance limits, portmanteau tests for white noise, and goodness-of-fit statistics) provided by standard softwares relying on the Box-Jenkins methodology can be quite misleading, as will be seen below. The consequences in terms of order selection, model validation, prediction, can be dramatic. It is therefore of great importance to develop tests and methods allowing to work with a broad class of ARMA models.

Section 2 of this paper deals with some preliminary definitions and interpretations. Examples of nonlinear processes admitting a linear representation are given in Section 3. Section 4 describes the asymptotic behaviour of the empirical autocovariances of a large class of stationary processes. We show how these results can be used for identifying the ARMA orders. In Section 5 we consider the estimation of the ARMA coefficients based on a least-squares criterion, with emphasis on the limiting behaviour of the estimators under mild dependence assumptions. In Section 6 we consider goodness-of-fit tests. Section 7 concludes. All proofs are collected in an appendix.

The main goal of this paper is to provide a review of recent results for ARMA models with uncorrelated but non independent errors. This paper also contains new results.

2. Basic definitions

In this paper, a second-order stationary real process $X = (X_t)_{t\in\mathbb{Z}}$, is said to admit an ARMA(p, q) representation, where p and q are integers, if for some constants $a_1, \ldots, a_p, b_1, \ldots, b_q$,

$$\forall t \in \mathbb{Z}, \quad X_t - \sum_{i=1}^{p} a_i X_{t-i} = \epsilon_t - \sum_{j=1}^{q} b_j \epsilon_{t-j}, \tag{12.1}$$

where $\epsilon = (\epsilon_t)$ is a white noise, that is, a sequence of centered uncorrelated variables, with common variance σ^2. It is convenient to write (12.1) as $\phi(L)X_t = \psi(L)\epsilon_t$, where L is the lag operator, $\phi(z) = 1 - \sum_{i=1}^{p} a_i z^i$ is the AR polynomial and $\psi(z) = 1 - \sum_{j=1}^{q} b_j z^j$ is the MA polynomial.

Different sub-classes of ARMA models can be distinguished depending on the noise assumptions. If in (12.1), ϵ is a strong white noise, that is a sequence of independent and identically distributed (iid) variables, then it is customary to call X a *strong* ARMA(p, q), and we will do this henceforth. If ϵ is a martingale difference (m.d.), that is a sequence such that $E(\epsilon_t|\epsilon_{t-1}, \epsilon_{t-2}, \ldots) = 0$, then X is called a *semi-strong* ARMA(p, q).

If no additional assumption is made on ϵ, that is if ϵ is only a weak white noise (not necessarily iid, nor a m.d.), X is called a *weak* ARMA(p, q). It is clear from these definitions that the following inclusions hold:

$$\{\text{strong ARMA}\} \subset \{\text{semi-strong ARMA}\} \subset \{\text{weak ARMA}\}.$$

It is often necessary to constrain the roots of the AR and MA polynomials. The following assumption is standard:

$$\phi(z)\psi(z) = 0 \Rightarrow |z| > 1. \tag{12.2}$$

Under (12.2): (i) a solution X_t of (12.1) is obtained as a linear combination of $\epsilon_t, \epsilon_{t-1}, \ldots$ (this solution is called non anticipative or causal), and (ii) ϵ_t can be written as a linear combination of $X_t, X_{t-1}, \ldots$ (equation (12.1) is said to be invertible). Moreover, it can be shown that if X satisfies (12.1) with $\phi(z)\psi(z) = 0 \Rightarrow |z| \neq 1$, then there exist a weak white noise ϵ^*, and some polynomials ϕ^* and ψ^*, with no root inside the unit disk, such that $\phi^*(B)X_t = \psi^*(B)\epsilon_t^*$. Therefore, (12.2) is not restrictive for weak ARMA models whose AR and MA polynomials have no root on the unit circle. The next example shows that this property does not hold for strong ARMA models.

Example 12.1 (Non causal strong AR(1)) Let the AR(1) model

$$X_t = aX_{t-1} + \eta_t, \quad \eta_t \text{ iid } (0, \sigma^2)$$

where $|a| > 1$. This equation has a stationary yet anticipative solution, of the form $X_t = -\sum_{i=1}^{\infty} a^{-i}\eta_{t+i}$, from which the autocovariance function of (X_t) is easily obtained as $\gamma_X(h) := \mathrm{Cov}(X_t, X_{t+h}) = \frac{\sigma^2}{a^{|h|}}\frac{1}{a^2-1}$. Now let $\epsilon_t^* = X_t - (1/a)X_{t-1}$. Then $E\epsilon_t^* = 0$, $\mathrm{Var}(\epsilon_t^*) = \sigma^2/a^2$, and it is straightforward to show that (ϵ_t^*) is uncorrelated. Therefore X admits the causal AR(1) representation $X_t = (1/a)X_{t-1} + \epsilon_t^*$. This AR(1) is not strong in general because, for example, it can be seen that $E\epsilon_t^* X_{t-1}^2 = E\eta_t^3/\{a^2(1 + a + a^2)\}$. Thus when $E\eta_t^3 \neq 0$, ϵ^* is not even a m.d.

The different assumptions on the noise have, of course, different consequences in terms of prediction. Assume that (12.2) holds. When the ARMA is semi-strong the optimal predictor is linear: that is, the best predictor of X_t in the mean-square sense is a linear function of its past values, and it can be obtained by inverting the MA polynomial. When the ARMA is weak, the same predictor is optimal among the linear functions of the past, and it is called optimal linear predictor, but not necessarily among all functions of the past.

The generality of the weak ARMA class can be assessed through the so-called Wold decomposition. Wold (1938) has shown that any purely non deterministic, second-order stationary process admits an infinite MA representation of the form

$$X_t = \epsilon_t + \sum_{i=1}^{\infty} c_i \epsilon_{t-i}, \tag{12.3}$$

where (ϵ_t) is the linear innovation process of X and $\sum_i c_i^2 < \infty$. Defining the MA(q) process $X_t(q) = \epsilon_t + \sum_{i=1}^{q} c_i \epsilon_{t-i}$, it is straightforward that

$$\|X_t(q) - X_t\|_2^2 = E\epsilon_t^2 \sum_{i>q} c_i^2 \to 0, \quad \text{when } q \to \infty.$$

Therefore any purely non deterministic, second-order stationary process is the limit, in the mean-square sense, of weak finite-order ARMA processes.

The principal conclusion of this section is that the linear model (12.3), which consists of the ARMA models and their limits, is very general under the noise uncorrelatedness, but can be restrictive if stronger assumptions are made.

3. Examples of exact weak ARMA representations

Many nonlinear processes admit ARMA representations. Examples of bilinear processes, Markov switching processes, threshold processes, deterministic processes, processes obtained by temporal aggregation of a strong ARMA, or by linear combination of independent ARMA processes can be found in Francq and Zakoïan (1998a, 2000). In this section, we will show that the powers of a GARCH(1,1) process (ϵ_t) admit an ARMA representation.

Consider the GARCH(1,1) model

$$\left\{ \begin{array}{rcl} \epsilon_t & = & \sqrt{h_t}\eta_t \\ h_t & = & \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \end{array} \right. \tag{12.4}$$

where α and β are nonnegative constants, ω is a positive constant, (η_t) is a strong white noise with unit variance. Under the constraint $\alpha + \beta < 1$, there exists a strictly stationary solution (ϵ_t), nonanticipative (i.e. such that ϵ_t is a function of the η_{t-i} with $i \geq 0$), and with a second order moment. This solution is clearly a semi-strong white noise. Since $E(\epsilon_t | \epsilon_{t-1}, \epsilon_{t-2}, \ldots) = h_t$, the innovation process of ϵ_t^2 is defined by $\nu_t = \epsilon_t^2 - h_t$. The second equation in (12.4) can be written as

$$\epsilon_t^2 = \omega + (\alpha + \beta)\epsilon_{t-1}^2 + \nu_t - \beta\nu_{t-1}. \tag{12.5}$$

This equation shows that, provided $E(\epsilon_t^4) < \infty$, the process (ϵ_t^2) admits an ARMA(1,1) representation. This representation is semi-strong because $E(\nu_t|\epsilon_{t-1}, \epsilon_{t-2}, \ldots) = 0$. The next theorem generalizes this well-known result.

THEOREM 12.1 *Suppose that (ϵ_t) is strictly stationary nonanticipative solution of Model (12.4) with $E\epsilon_t^{4m} < \infty$, where m is an integer. Then (ϵ_t^{2m}) admits a weak ARMA(m, m) representation.*

REMARK 12.1 A similar theorem was proved in the framework of Markov-switching GARCH(p, q) models by Francq and Zakoïan (2003). This class encompasses the standard GARCH model so the theorem of the present paper can be seen as a corollary of a more general theorem. However, the result was obtained for the general class with a more complex proof requiring vector representations. The proof presented in the appendix gives an alternative (simpler) way to show the existence of the ARMA representations in the GARCH(1,1) case.

REMARK 12.2 As can be seen in the appendix, the ARMA coefficients can be derived from the coefficients of model (12.4) and from moments of η_t.

REMARK 12.3 Contrary to the representation (12.5), the ARMA representations for $m > 1$ are not semi-strong in general. The noises involved in these representations are not m.d.'s. This can be seen from tedious computations involving numerical calculations.

4. Sample ACVF and related statistics

In time series analysis, the autocovariance function (ACVF) is an important tool because it provides the whole linear structure of the series. More precisely, it is easy to compute the ACVF from the Wold representation of a regular stationary process. Conversely, the knowledge of the ACVF is sufficient to determine the ARMA representation (when existing).

Consider observations $X_1, \ldots, X_n$ of a centered second-order stationary process. The sample ACVF and autocorrelation function (ACF) are defined, respectively, by

$$\hat{\gamma}(h) = \hat{\gamma}(-h) = \frac{1}{n} \sum_{t=1}^{n-h} X_t X_{t+h}, \quad \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad 0 \leq h < n.$$

We know that for strictly stationary ergodic processes, $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are strongly consistent estimators of the theoretical autocovariance $\gamma(h)$

and the theoretical autocorrelation $\rho(h)$. For $0 \le m < n$, let $\hat{\gamma}_{0:m} := \{\hat{\gamma}(0), \hat{\gamma}(1), \dots, \hat{\gamma}(m)\}'$ and $\gamma_{0:m} := \{\gamma(0), \gamma(1), \dots, \gamma(m)\}'$. Similarly, denote by $\hat{\rho}_{1:m}$ and $\rho_{1:m}$ the vectors of the first m sample autocorrelations and theoretical autocorrelations.

4.1 Asymptotic behaviour of the sample ACVF

Extending the concept of second-order stationarity, a real-valued process (X_t) is said to be stationary up to the order k if $E|X_t|^k < \infty$ for all t, and $EX_{t_1}X_{t_2}\dots X_{t_k} = EX_{t_1+h}X_{t_2+h}\dots X_{t_k+h}$ for all $k \in \{1, 2, \dots, m\}$ all $h, t_1, \dots, t_k \in \mathbb{Z}$.

THEOREM 12.2 *If (X_t) is centered and stationary up to the order 4, and if*

$$\sum_{\ell=-\infty}^{+\infty} |\sigma_{h,k}(\ell)| < \infty, \quad \textit{for all } k, h \in \{0, 1 \dots, m\}$$

where $\sigma_{h,k}(\ell) = Cov(X_t X_{t+h}, X_{t+\ell}X_{t+\ell+k})$, then

$$\lim_{n\to\infty} n\,Var(\hat{\gamma}_{0:m}) = \Sigma_{\hat{\gamma}_{0:m}}, \quad \textit{where } \Sigma_{\hat{\gamma}_{0:m}} = \left(\sum_{\ell=-\infty}^{+\infty} \sigma_{h,k}(\ell)\right)_{0\le h,k\le m}.$$

REMARK 12.4 We have $\Sigma_{\hat{\gamma}_{0:m}} = \sum_{\ell=-\infty}^{+\infty} \mathrm{Cov}(\Upsilon_t, \Upsilon_{t+\ell})$ where $\Upsilon_t := X_t(X_t, X_{t+1}, \dots, X_{t+m})'$. Thus $(2\pi)^{-1}\Sigma_{\hat{\gamma}_{0:m}}$ can be interpreted as the spectral density of (Υ_t) at the zero frequency.

To show the asymptotic normality of the sample ACV's, additional moment conditions and weak dependence conditions are needed. Denote by $(\alpha_X(k))_{k\in N^*}$ the sequence of strong mixing coefficients of any process $(X_t)_{t\in\mathbb{Z}}$. Romano and Thombs (1996) showed that under the assumption

A1: $\sup_t E|X_t|^{4+2\nu} < \infty$ and $\sum_{k=0}^{\infty}\{\alpha_X(k)\}^{\nu/(2+\nu)} < \infty$

for some $\nu > 0$, $\sqrt{n}\,(\hat{\gamma}_{0:m} - \gamma_{0:m})$ is asymptotically normal with mean 0 and variance $\Sigma_{\hat{\gamma}_{0:m}}$. The mixing condition in **A1** is valid for large classes of processes (see Pham (1986), Carrasco and Chen (2002)) for which the (stronger) condition of β-mixing with exponential decay can be shown. If the mixing coefficients tend to zero at an exponential rate, ν can be chosen arbitrarily small. The moment condition is mild since $EX_t^4 < \infty$ is required for the existence of $\Sigma_{\hat{\gamma}_{0:m}}$. Romano and Thombs (1996) proposed resampling methods to obtain confidence intervals and to perform tests on the ACVF and ACF. Berlinet and Francq (1997, 1999) employed HAC type estimators to estimate $\Sigma_{\hat{\gamma}_{0:m}}$ (see *e.g.* Andrews (1991) for details about HAC estimators). In view of Remark 12.4, alternative estimators of $\Sigma_{\hat{\gamma}_{0:m}}$ can be obtained by smoothing the periodogram of Υ_t at the zero frequency (see *e.g.* Brockwell and Davis, 1991, Sections 10.4 and 11.7) or by using an AR spectral density estimator (see *e.g.* Brockwell and Davis, 1991, Section 10.6, and Berk, 1974). In this paper we use the AR spectral density estimator, studied in Francq, Roy and Zakoïan (2003), and defined by

$$\hat{\Sigma}_{\hat{\gamma}_{0:m}} = \hat{\mathcal{A}}_r(1)^{-1}\hat{\Sigma}_r\hat{\mathcal{A}}_r(1)^{-1}, \tag{12.6}$$

where $\hat{\mathcal{A}}_r(\cdot)$ and $\hat{\Sigma}_r$ denote the fitted AR polynomial and the variance of the residuals in the multivariate regression of Υ_t on $\Upsilon_{t-1}, \ldots, \Upsilon_{t-r}$, for $t = r+1, \ldots, n-m$. The AR polynomial is fitted using the Durbin-Levinson algorithm. The order r is selected minimizing the BIC information criteria.

Using the delta method, it is easy to obtain the asymptotic distribution of differentiable functions of $\hat{\gamma}_{0:m}$. In particular, we have

$$\sqrt{n}\left(\hat{\rho}_{1:m} - \rho_{1:m}\right) \Rightarrow \mathcal{N}\left(0, \Sigma_{\hat{\rho}_{1:m}}\right), \quad \Sigma_{\hat{\rho}_{1:m}} = J_m \Sigma_{\hat{\gamma}_{0:m}} J_m' \tag{12.7}$$

and $J_m = \partial\rho_{1:m}/\partial\gamma_{0:m}'$. For strong ARMA processes, Matrix $\Sigma_{\hat{\rho}_{1:m}}$ is given by the so-called Bartlett's formula, which involves the theoretical autocorrelation function $\rho(\cdot)$ only. In particular, for an iid sequence, $\hat{\rho}(1), \ldots, \hat{\rho}(h)$ are approximately independent and $\mathcal{N}(0, n^{-1})$ distributed, for large n. This can be used to detect whether a given series is an iid sequence or not. For a strong MA(q), $\sqrt{n}\hat{\rho}(h)$ is asymptotically normal with mean 0 and variance $\sum_{k=-q}^{q} \rho^2(k)$, for any $h > q$. Replacing $\rho(k)$ by $\hat{\rho}(k)$, this can be used to select the order of a strong MA.

Unfortunately, in the case of weak MA or ARMA models, $\Sigma_{\hat{\rho}_{1:m}}$ is much more complicated, and involves not only $\rho(\cdot)$, but also fourth order cumulants.

4.2 Misuse of the Bartlett formula

The usual text book formula for $\Sigma_{\hat{\rho}_{1:m}}$, the above-mentioned Bartlett's formula, is based on strong assumptions (see Mélard and Roy, 1987, Mélard, Paesmans and Roy, 1991, and Roy,1989). In particular, the Bartlett formula is not valid when the linear innovations are not iid. It can be shown that the value given by the Bartlett formula can deviate markedly from the true value of $\Sigma_{\hat{\rho}_{0:m}}$. As can be seen in the following example, this result is essential in the identification stage of the Box-Jenkins methodology.

Example 12.2 Figure 12.1 displays the sample ACF of a simulation of length 5000 of the GARCH(1,1) model (12.4) where (η_t) is i.i.d $\mathcal{N}(0, 1)$,

$\omega = 1$, $\alpha = 0.3$ and $\beta = 0.55$. The solution of (12.4) is a white noise with fourth-order moments, but the ϵ_t's are not independent (the ϵ_t^2's are correlated). The Bartlett formula applied to a strong white noise gives $\Sigma_{\hat{\rho}_{1:m}} = I_m$. We deduce that, asymptotically, the sample autocorrelations of a strong white noise are between the bounds $\pm 1.96/\sqrt{n}$ with probability 0.95. The usual procedure consists in rejecting the white noise assumption and fitting a more complex ARMA model when numerous sample autocorrelations lie outside the bounds $\pm 1.96/\sqrt{n}$ (or when one sample autocorrelation is quite outside the bounds). This procedure is absolutely valid for testing the independence assumption, but may lead to incorrect conclusions when applied for testing the weak white noise assumption. In the left graph of Figure 12.1, it can be seen that numerous sample autocorrelations lie outside the bounds $\pm 1.96/\sqrt{n}$. The sample autocorrelations at lag 2 and 3 are much less that $-1.96/\sqrt{n}$. This leads us to reject the strong white noise assumption. However, this does not mean that the weak white noise must be rejected, *i.e.* that some sample autocorrelations are significant. Actually, all the sample autocorrelations are inside the true significance limits (in thick dotted lines). The true 5% significance limits for the sample ACF are of course unknown in practice. In the right graph of Figure 12.1, an estimate (based on the AR spectral density estimator of $\Sigma_{\hat{\rho}_{1:m}}$ described in Section 4.1) of these significance limits is plotted. From the right graph, there is no evidence against the weak white noise assumption (but from the left graph the strong white noise model is rejected). We draw the conclude that, to detect weak ARMA models, it is important to use more robust estimators of $\Sigma_{\hat{\rho}_{1:m}}$ than those given by the Bartlett formula.

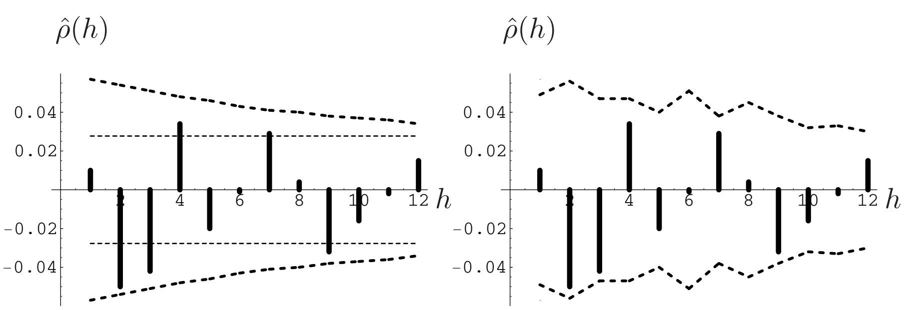

Figure 12.1. Sample ACF of a simulation of length $n = 5000$ of the weak (semi-strong) white noise (12.4). In the left graph, the horizontal dotted lines $\pm 1.96/\sqrt{n}$ correspond to the asymptotic 5% significance limits for the sample autocorrelations of a strong white noise. The thick dotted lines about zero represent the asymptotic 5% significance limits for the sample autocorrelations of ϵ_t. In the right graph, the thick dotted lines about zero represent an estimate of the 5% significance limits for the sample autocorrelations of ϵ_t.

4.3 Behaviour of the sample partial autocorrelations

The partial autocovariance function (PACF), denoted by $r(\cdot)$, is a very convenient tool to detect AR(p) processes. Since $r(h)$ is a differentiable function of $\rho(1), \rho(2), \ldots, \rho(h)$, the asymptotic behaviour of the sample PACF $\hat{r}(\cdot)$ is given by a formula similar to (12.7).

It is known that, for strong AR(p) processes, $\sqrt{n}\hat{r}(h)$ is asymptotically $\mathcal{N}(0,1)$ distributed, for each $h > p$. This is very convenient to detect the order of strong AR processes. Of course, the limiting distribution is no more $\mathcal{N}(0,1)$ for weak AR models.

The following result shows that a comparison between the sample ACF and sample PACF can be useful to detect a weak white noise.

Theorem 12.3 *If (X_t) is a stationary weak white noise satisfying* **A1**, *then*

$$\hat{\rho}(h) - \hat{r}(h) = O_P(n^{-1}). \tag{12.8}$$

Remark 12.5 This result indicates that, for a weak white noise, the sample ACF and sample PACF should be very close. To our knowledge, however, no formal test based on this principle exists. Theorem 12.3 also entails that, for a weak white noise, the asymptotic distributions of $\sqrt{n}\hat{\rho}_{1:m}$ and $\sqrt{n}\hat{r}_{1:m} := \sqrt{n}\,\{\hat{r}(1), \ldots, \hat{r}(m)\}$ are the same.

4.4 Portmanteau tests for weak white noise assumption

It is possible to test the assumption $H_0 : \rho(1) = \cdots = \rho(m) = 0$ by a so-called portmanteau test. The initial version of the test is based on the Box and Pierce (1970) statistic $Q_m = n\sum_{h=1}^{m}\hat{\rho}^2(h)$. Under the assumption H_0' that (X_t) is a strong white noise, we know from the Bartlett formula that the asymptotic distribution of $\sqrt{n}\hat{\rho}_{1:m}$ is $\mathcal{N}(0, I_m)$. Therefore the asymptotic distribution of Q_m is χ^2_m under H_0'. The standard test procedure consists in rejecting the null hypothesis H_0 if $Q_m > \chi^2_m(1-\alpha)$, where $\chi^2_m(1-\alpha)$ denotes the $(1-\alpha)$-quantile of a χ^2 distribution with m degrees of freedom. Ljung and Box (1978) argued that, in the case of a Gaussian white noise, the finite sample distribution of the test statistic is more accurately approximated by a χ^2_m when Q_m is replaced by $\tilde{Q}_m = n(n+2)\sum_{h=1}^{m}\frac{\hat{\rho}^2(h)}{n-h}$. Nowadays, all the time series packages incorporate the Ljung-Box version of the portmanteau test.

It is easy to see that the asymptotic distribution of Q_m (and $\tilde{Q}_m$) can be very different from the χ^2_m when H_0 is satisfied, but H_0' is not (see Lobato (2001) and Francq, Roy and Zakoïan (2003)). Therefore the standard portmanteau test is rather an independence test than an uncorrelatedness test. Thus, when the standard portmanteau test leads to the rejection, *i.e.* $Q_m > \chi^2_m(1-\alpha)$, the practitioner does not know whether (i) H_0 is actually false, or (ii) H_0 is true but H_0' is false. The problem can be solved by using the estimators of $\Sigma_{\hat{\gamma}_{0:m}}$ mentioned in Section 4.1.

THEOREM 12.4 *Let (X_t) be a stationary process and let $\Upsilon_t := (X_t^2, X_tX_{t+1}, \ldots, X_tX_{t+m})$. Assume that $Var(\Upsilon_t)$ is non-singular. If (X_t) is a weak white noise satisfying* **A1** *and if $\hat{\Sigma}_{\hat{\rho}_{1:m}}$ is a weakly consistent estimator of $\Sigma_{\hat{\rho}_{1:m}}$, then the limit law of the statistics*

$$Q_m^{\rho} = n\hat{\rho}_{1:m}'\hat{\Sigma}^{-1}_{\hat{\rho}_{1:m}}\hat{\rho}_{1:m} \quad \textit{and} \quad Q_m^{r} = n\hat{r}_{1:m}'\hat{\Sigma}^{-1}_{\hat{\rho}_{1:m}}\hat{r}_{1:m} \tag{12.9}$$

is the χ^2_m distribution.

REMARK 12.6 Replacing Q_m by Q_m^{ρ} in the standard portmanteau test for strong white noise, we obtain a portmanteau test for weak white noise. Another portmanteau test for weak white noise, based on the sample PACF, is obtained when the statistic Q_m^r is used. Monte Carlo experiments reveal that the portmanteau test based on the sample PACF may be more powerful than that based on the sample ACF. Intuitively, this is the case for alternatives such that the magnitude of the PACF is larger than that of the ACF (in particular, for MA models).

EXAMPLE 12.3 *(Example 12.2 continued)* The sample PACF is displayed in Figure 12.2. It is seen that this PACF is very close to the sample ACF displayed in the right graph of Figure 12.1. This is not surprising in view of Theorem 12.3. The results of the standard portmanteau tests are not presented but lead to reject the strong white noise assumption ($P(\chi^2_m > Q_m)$ and $P(\chi^2_m > \tilde{Q}_m)$ are less than 0.001 for $m > 1$). Figure 12.3 show that the weak white noise assumption is not rejected by the portmanteau tests based on the statistic Q_m^{ρ} (the portmanteau tests based on the PACF lead to similar results).

4.5 Application to the identification stage

To detect the order of a MA, a common practice is to draw the sample ACF, as in Figure 12.1. If for some not too large integer q_0, almost all the $\hat{\rho}(h)$ such that $h > q_0$ are inside the bounds $\pm 1.96 n^{-1/2}\left\{\sum_{j=-h+1}^{j=h-1}\hat{\rho}^2(j)\right\}^{1/2}$, practitioners consider that a MA(q_0) is a reasonable model. Although meaningful for detecting the order of a strong MA process, this procedure is not valid for identifying weak MA models. Indeed, we have seen that the bounds given by the Bartlett formula are not correct in

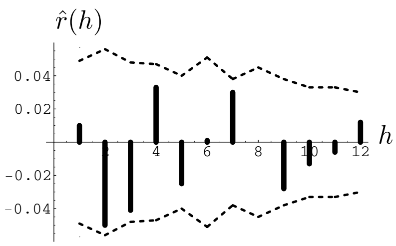

Figure 12.2. As for the right graph in Figure 12.1, but for the Sample PACF.

m	1	2	3	4
Q_m^ρ	0.2	3.3	6.2	8.3
$P(\chi_m^2 > Q_m^\rho)$	0.696	0.197	0.101	0.081
m	7	8	9	10
m	7	8	9	10
Q_m^ρ	11.5	11.5	14.3	15.2
$P(\chi_m^2 > Q_m^\rho)$	0.120	0.176	0.113	0.123

Figure 12.3. Portmanteau tests of the weak white noise assumption (based on (12.9)) for a simulation of length $n = 5000$ of Model (12.4).

this framework. A procedure which remains valid for weak models, is obtain by replacing $\sum_{j=-h+1}^{j=h-1} \hat{\rho}^2(j)$ by the h-th diagonal term of the matrix $\hat{\Sigma}_{\hat{\rho}_{1:m}}$ given in (12.7) and (12.6). With this modification, the identification of (weak) MA models can be done in the usual way.

Similarly, the asymptotic distribution of $\hat{r}_{1:m}$, given in Section 4.3 can be used to detect the order of an AR model. For mixed ARMA processes, different statistics have been introduced in the literature (*e.g.* the corner method (Béguin, Gouriéroux and Monfort, 1980), the epsilon algorithm (Berlinet, 1984), Glasbey's generalized autocorrelations (1982)). Initially the methods based on these statistics were developed for strong ARMA models, using the conventional Bartlett formula. Without any modification, their application to weak ARMA models can misguide the analyst and result in an inappropriate model being selected. However, all the above-mentioned statistics are differentiable functions of $\hat{\rho}_{1:m}$. Thus their asymptotic distributions can be derived from (12.7). Using this correction, these methods remain valid for selecting weak ARMA models.

Other popular approaches for identifying the ARMA orders are founded on information criteria such as AIC and BIC. It was shown in Francq and Zakoïan (1998b) that, asymptotically, the orders are not underestimated when these criteria are applied to weak ARMA models. The consistency of the BIC criterion is an open issue in the weak ARMA framework.

5. Estimation

Following the seminal papers by Mann and Wald (1943) and Walker (1964), numerous works have been devoted to the estimation of ARMA models. Hannan (1973) showed the strong consistency of several estimators, under the ergodicity assumption for the observed process, and the asymptotic normality under the m.d. assumption for the noise. Dunsmuir and Hannan (1976) and Rissanen and Caines (1979) extended these results to multivariate processes. Tiao and Tsay (1983) considered a nonstationary case with an iid noise. Mikosch, Gadrich, Klüppelberg and Adler (1995) considered ARMA with infinite variance and iid noise. Kokoszka and Taqqu (1996) considered fractional ARIMA with infinite variance and iid noise. Following Dickey and Fuller (1979), the estimation of ARMA models with unit roots have been extensively investigated. All these authors assumed (at least for the asymptotic normality) that the linear innovations constitute a m.d. This assumption is not satisfied for ARMA representations of non linear processes.

We will investigate the asymptotic behaviour of the ordinary least squares (OLS) estimator when the linear innovation process is not a m.d. Alternative estimators, such as those obtained from the generalized method of moments (see Hansen (1982), have been proposed (see Broze, Francq and Zakoïan (2002)). Denote by $\theta_0 = (a_1, \dots, a_p, b_1, \dots, b_q)'$ the true value of the unknown parameter.

5.1 Consistency

The assumptions are the following.

A2: X is strictly stationary and ergodic.

A3: 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.

A4: $p + q > 0$ and $a_p^2 + b_q^2 \neq 0$ (by convention $a_0 = b_0 = 1$).

A5: $\sigma^2 > 0$.

A6: $\theta_0 \in \Theta^*$ where Θ^* is a compact subspace of the parameter space

$$\begin{aligned}\Theta := \big\{ & \theta = (\theta_1, \dots, \theta_{p+q}) \in \mathbb{R}^{p+q} : \ \phi_\theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p \\ & \text{and } \psi_\theta(z) = 1 - \theta_{p+1} z - \cdots - \theta_{p+q} z^q \\ & \text{have all their zeros outside the unit disk} \big\}.\end{aligned}$$

Note that **A2** is equivalent to **A2'**: ϵ is strictly stationary and ergodic.

For all $\theta \in \Theta$, let

$$\epsilon_t(\theta) = \psi_\theta^{-1}(B)\phi_\theta(B)X_t = X_t + \sum_{i=1}^{\infty} c_i(\theta) X_{t-i}.$$

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

$$e_t(\theta) = X_t - \sum_{i=1}^{p} \theta_i X_{t-i} + \sum_{i=1}^{q} \theta_{p+i} e_{t-i}(\theta) \qquad (12.10)$$

where the unknown starting values are set to zero: $e_0(\theta) = e_{-1}(\theta) = \ldots = e_{-q+1}(\theta) = X_0 = X_{-1} = \ldots = X_{-p+1} = 0$. The random variable $\hat{\theta}_n$ is called least squares (LS) estimator if it satisfies, almost surely,

$$Q_n(\hat{\theta}_n) = \min_{\theta \in \Theta^*} Q_n(\theta), \qquad Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} e_t^2(\theta). \qquad (12.11)$$

Hannan (1973) showed the following result using the spectral analysis. The proof we give in the appendix is quite different.

THEOREM 12.5 *Under Assumptions* **A2**-**A6**, $\hat{\theta}_n \longrightarrow \theta_0$ *almost surely as* $n \to \infty$.

5.2 Asymptotic normality

The asymptotic normality is established in Francq and Zakoïan (1998a), using the additional assumptions **A1** and

A6 : θ belongs to the interior of Θ^*.

Assumption **A1** can be replaced by

A1' : $E|\epsilon_t|^{4+2\nu} < \infty$ and $\sum_{k=0}^{\infty} [\alpha_\epsilon(k)]^{\frac{\nu}{2+\nu}} < \infty$ for some $\nu > 0$.

Note that Assumptions **A1** and **A1'** are not equivalent.

THEOREM 12.6 *Under Assumption* **A1** *or* **A1'** *and Assumptions* **A2**-**A6**,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{d}{\rightsquigarrow} \mathcal{N}(0, J^{-1} I J^{-1}) \text{ as } n \to \infty, \qquad (12.12)$$

where $I = I(\theta_0)$, $J = J(\theta_0)$,

$$I(\theta) = \frac{1}{4} \lim_{n\to\infty} var \left\{ \sqrt{n} \frac{\partial}{\partial\theta} Q_n(\theta) \right\}, \quad J(\theta) = \frac{1}{2} \lim_{n\to\infty} \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta) \quad a.s.$$

REMARK 12.7 Matrices I and J can be computed explicitly. Indeed, from (12.1) we have

$$-X_{t-i} = \psi(B) \frac{\partial}{\partial\theta_i} \epsilon_t \quad \text{and} \quad 0 = -\epsilon_{t-j} + \psi(B) \frac{\partial}{\partial\theta_{p+j}} \epsilon_t$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Thus

$$\partial\epsilon_t/\partial\theta = (u_{t-1}, \ldots, u_{t-p}, v_{t-1}, \ldots, v_{t-q})'$$

where

$$u_t = -\phi^{-1}(B)\epsilon_t := -\sum_{i=0}^{\infty} \phi_i^* \epsilon_{t-i} \quad \text{and} \quad v_t = \psi^{-1}(B)\epsilon_t := \sum_{i=0}^{\infty} \psi_i^* \epsilon_{t-i}.$$

Hence

$$\begin{aligned} J &= \operatorname{Var}\left(\frac{\partial}{\partial\theta}\epsilon_t\right)_{\theta=\theta_0} = \operatorname{Var}(u_{t-1}, \ldots, u_{t-p}, v_{t-1}, \ldots, v_{t-q})' \\ &= \operatorname{Var}\sum_{i=1}^{\infty} \epsilon_{t-i}\lambda_i = \sigma^2 \Lambda_\infty \Lambda_\infty' \end{aligned}$$

and

$$\begin{aligned} I &= \sum_{h=-\infty}^{+\infty} \operatorname{Cov}\left\{\epsilon_t \frac{\partial}{\partial\theta}\epsilon_t, \epsilon_{t+h}\frac{\partial}{\partial\theta}\epsilon_{t+h}\right\} \\ &= \sum_{i,j=1}^{\infty} \lambda_i \sum_{h=-\infty}^{+\infty} \operatorname{Cov}(\epsilon_t\epsilon_{t-i}, \epsilon_{t+h}\epsilon_{t+h-j})\,\lambda_j' = \Lambda_\infty \Sigma_{\hat{\gamma}^\epsilon_{1:\infty}} \Lambda_\infty' \end{aligned}$$

where, for $m = 1, 2, \ldots, \infty$, the matrix $\Sigma_{\hat{\gamma}^\epsilon_{1:m}}$ is the asymptotic variance of $\sqrt{n}\,\{\hat{\gamma}_\epsilon(1), \ldots, \hat{\gamma}_\epsilon(m)\}'$, and

$$\Lambda_m = (\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_m) = \begin{pmatrix} -1 & -\phi_1^* & \cdots & & & & -\phi_{m-1}^* \\ 0 & -1 & & & & & \vdots \\ \vdots & & & & & & \\ 0 & & & -1 & -\phi_1^* & \cdots & -\phi_{m-p}^* \\ 1 & \psi_1^* & \cdots & & & & \psi_{m-1}^* \\ 0 & 1 & & & & & \vdots \\ \vdots & & & & & & \\ 0 & & & 1 & \psi_1^* & \cdots & \psi_{m-q}^* \end{pmatrix}. \qquad (12.13)$$

5.3 Asymptotic accuracy of LSE for weak and strong ARMA

In the strong ARMA case, we have

$$I = \frac{1}{4}\operatorname{var}\left\{\frac{\partial}{\partial\theta}\epsilon_t^2(\theta_0)\right\} = \operatorname{var}\left\{\epsilon_t \frac{\partial}{\partial\theta}\epsilon_t(\theta_0)\right\} = \sigma^2 J.$$

The asymptotic covariance matrix of the LSE is then $S = \sigma^2 J^{-1}$. In the weak ARMA case, the true asymptotic variance Σ can be very different from S (see Example 12.4 below) For the statistical inference on the parameter, in particular the Student's t significant test, the standard time series analysis softwares use empirical estimators of S. Under ergodicity and moment assumptions, these estimators converges to S, but, in general, they do not converge to the true asymptotic variance Σ. This can of course be a serious cause of mistake in model selection. Francq and Zakoïan (2000) have proposed an estimator of Σ which is consistent for both weak and strong ARMA cases.

Example 12.4 *(Example 12.3 continued)* Consider the AR(1) model $X_t - aX_{t-1} = \epsilon_t$ where ϵ_t is the weak white noise defined by the GARCH(1,1) equation (12.4). The LSE of a is simply $\hat{a} = \sum_{t=2}^{n} X_t X_{t-1} / \sum_{t=2}^{n} X_{t-1}^2$. From Theorem 12.6 we know that $n^{1/2}(\hat{a} - a)$ converges in law to the $\mathcal{N}(0, \Sigma)$ distribution, where Σ depends on a, ω, α and β. It is also well known that, when (ϵ_t) is a strong white noise (*i.e.* when $\alpha = \beta = 0$), $n^{1/2}(\hat{a} - a)$ converges in law to the $\mathcal{N}(0, S)$ distribution, where $S = 1 - a^2$. Figure 12.4 displays the ratio Σ/S for several values of a, α and β. It can be seen that Σ and S can be very different.

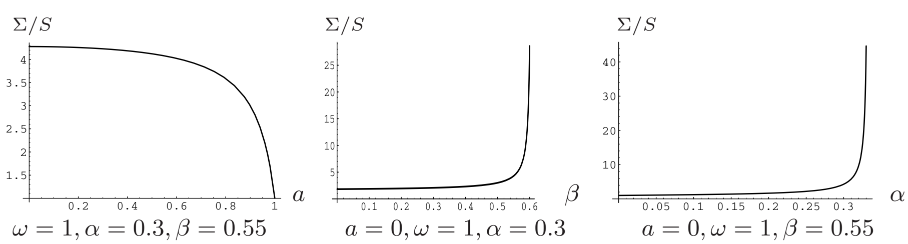

Figure 12.4. Comparison of the true asymptotic covariance Σ and $S = 1 - a^2$ for the LS estimator of a in the model $X_t = aX_{t-1} + \epsilon_t$, where (ϵ_t) follows the GARCH(1,1) equation (12.4).

6. Validation

The goodness of fit of a given ARMA(p, q) model is often judged by studying the residuals $\hat{\epsilon} = (\hat{\epsilon}_t)$. If the model is appropriate, then the residual autocorrelations should not be too large. Box and Pierce (1970) gave the asymptotic distribution of the residual autocorrelations of strong ARMA models. Recently, Francq, Roy and Zakoïan (2003) considered the weak ARMA case. We summarize their results. Let $\hat{\rho}_{1:m}^{\hat{\epsilon}}$ be the vector of the first m residual autocorrelations.

We use the notations of Remark 12.7 and the additional notation

$$\Sigma_{\hat{\gamma}^{\epsilon}_{1:m\times 1:m'}} = \lim_{n\to\infty} n\mathrm{Cov}\left[\{\hat{\gamma}_\epsilon(1),\ldots,\hat{\gamma}_\epsilon(m)\}', \{\hat{\gamma}_\epsilon(1),\ldots,\hat{\gamma}_\epsilon(m')\}'\right].$$

Under the assumptions of Theorem 12.6, it can be shown that $\sqrt{n}\hat{\rho}^{\hat{\epsilon}}_{1:m} \overset{d}{\leadsto} \mathcal{N}(0, \Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}})$, where

$$\begin{aligned}\Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}} &= \sigma^{-4}\Sigma_{\hat{\gamma}^{\epsilon}_{1:m}} + \Lambda'_m J^{-1} I J^{-1}\Lambda_m - \sigma^{-2}\Lambda'_m J^{-1}\Lambda_\infty \Sigma_{\hat{\gamma}^{\epsilon}_{1:\infty\times 1:m}} \\ &\quad -\sigma^{-2}\Sigma_{\hat{\gamma}^{\epsilon}_{1:m\times 1:\infty}}\Lambda'_\infty J^{-1}\Lambda_m. \end{aligned} \tag{12.14}$$

Note that in the strong ARMA case, $\Sigma_{\hat{\gamma}^{\epsilon}_{1:m}} = \sigma^4 I_m$ and

$$\Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}} = I_m - \Lambda'_m\{\Lambda_\infty\Lambda'_\infty\}^{-1}\Lambda_m \simeq I_m - \Lambda'_m\{\Lambda_m\Lambda'_m\}^{-1}\Lambda_m$$

is close to a projection matrix with $m-(p+q)$ eigenvalues equal to 1, and $(p+q)$ eigenvalues equal to 0, and we retrieve the result given in Box and Pierce (1970). Matrices Λ_m only depend on the ARMA parameters and can be estimated by a plug-in approach. Matrices $\Sigma_{\hat{\gamma}^{\epsilon}_{1:m}}$ can be estimated by the methods described in Section 4.1. This leads to a consistent estimator of $\Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}}$, which provides estimated significance limits for the sample autocorrelations. This can be used in a graph, similar to that of the right panel of Figure 12.1, to judged whether the selected weak ARMA(p,q) model is adequate or not.

Other very popular diagnostic checking tools are the portmanteau tests based on the residual autocorrelations. Let $Q^{\rho_{\hat{\epsilon}}}_m := n\sum_{i=1}^m \hat{\rho}^2_{\hat{\epsilon}}(i)$ be the Box-Pierce portmanteau statistic based on the first m sample autocorrelations. The portmanteau statistic for residuals is not defined as in (12.9), because $\Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}}$ can be singular. From (12.13), it is easy to see that the statistics $Q^{\rho_{\hat{\epsilon}}}_m$ converges in distribution, as $n\to\infty$, to

$$\mathbf{Z}_m(\xi_m) := \sum_{i=1}^m \xi_{i,m} Z_i^2 \tag{12.15}$$

where $\xi_m = (\xi_{1,m},\ldots,\xi_{m,m})'$ is the eigenvalues vector of $\Sigma_{\hat{\rho}^{\hat{\epsilon}}_{1:m}}$, and $Z_1,\ldots,Z_m$ are independent $\mathcal{N}(0,1)$ variables. The distribution of the quadratic form $\mathbf{Z}_m(\xi_m)$ can be computed using the algorithm by Imhof (1961).

Example 12.5 *(Example 12.4 continued)* We simulated a realization of length $n = 5000$ of the AR(1) model $X_t - aX_{t-1} = \epsilon_t$ with $a = 0.5$ and where (ϵ_t) satisfies the GARCH(1,1) equation (12.4). The LS estimate of a is $\hat{a} = 0.483$. Figure 12.5 displays the sample autocorrelations,

$\hat{\rho}_{\hat{\epsilon}}(h)$, $h = 1, \ldots, 12$, of the residuals $\hat{\epsilon}_t = X_t - \hat{a}X_{t-1}$, $t = 2, \ldots, n$. For a strong AR(1) model, we know that $\hat{\rho}_{\hat{\epsilon}}(h)$ should lie between the bounds $\pm 1.96 n^{-1/2} \left\{1 - a^{2h-2}(1 - a^2)\right\}$ with probability close to 95%. These 5% significance limits are plotted in thin dotted lines. It can be seen that numerous residual autocorrelations are close or outside the thin dotted lines, which can be considered as an evidence against the strong AR(1) model. This is confirmed by the standard portmanteau tests given in Figure 12.6. For the simulated weak AR(1), the correct 5% significance limits are not the thin dotted lines. Estimates of the true 5% significance limits are plotted in thick dotted lines. Since all the residual autocorrelations are inside these limits, there is no evidence against the weak AR(1) model. This is confirmed by the portmanteau tests (based on the statistic $Q_m^{\rho_{\hat{\epsilon}}}$ and on the limiting distribution (12.15)), presented in Figure 12.7.

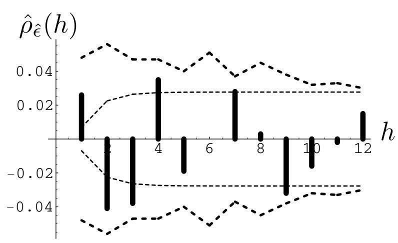

Figure 12.5. Residual ACF of a simulation of the model of Example 12.5

m	1	2	3	12
$P(\chi^2_m > Q^{\rho}_m)$	n.d.	0.001	0.000	0.000

Figure 12.6. Portmanteau test for a strong AR(1)

m	1	2	3	12
p-value	0.295	0.187	0.125	0.185

Figure 12.7. Portmanteau test for a weak AR(1)

7. Conclusion

The principal thrust of this paper is to point out that : (i) the standard Box-Jenkins methodology can lead to erroneous conclusions when the independence assumption on the noise is in failure; (ii) this methodology can be accommodated to handle weak ARMA models. We made a survey of the recent literature on this topic and also presented some complementary new results and illustrations. Of course, many issues remain uncovered, such as efficiency of the inference procedures, testing the different classes of ARMA models, extension to the multivariate framework.

Acknowledgments Part of this material has been presented in a series of seminars organized by R. Roy at the University of Montréal. We would like to thank the participants of these seminars. The authors are grateful to the referee for his comments.

Appendix: Proofs

Proof of Theorem 12.1. Let for $i \geq j \geq 0$,

$$\xi_{i,j} = \binom{i}{j} \omega^{i-j} E(\alpha\eta_t^2 + \beta)^j \frac{\mu_{2i}}{\mu_{2j}} \quad \text{where} \quad \mu_{2i} = E\eta_t^{2i}.$$

We have, for $i = 0, \ldots, 2m$

$$E\epsilon_t^{2i} = \mu_{2i} \sum_{j=0}^{i} \binom{i}{j} \omega^{i-j} E(\alpha\eta_{t-1}^2 + \beta)^j E(h_{t-1}^j) = \sum_{j=0}^{i} \xi_{i,j} E(\epsilon_{t-1}^{2j}), \qquad \text{(A.1)}$$

using the independence between η_t and h_t. Similarly, for $k > 1$ and $i = 0, \ldots, m$

$$E(\epsilon_t^{2i}\epsilon_{t-k}^{2m}) = \sum_{j=0}^{i} \xi_{i,j} E(\epsilon_{t-1}^{2j}\epsilon_{t-k}^{2m}).$$

It follows that for $k > 1$ and $i = 0, \ldots, m$

$$\text{Cov}(\epsilon_t^{2i}, \epsilon_{t-k}^{2m}) = \sum_{j=1}^{i} \xi_{i,j} \text{Cov}(\epsilon_{t-1}^{2j}, \epsilon_{t-k}^{2m}). \qquad \text{(A.2)}$$

Denoting by L the lag operator and writing, for any bivariate stationary process (X_t, Y_t), $L\text{Cov}(X_t, Y_t) = \text{Cov}(X_{t-1}, Y_t)$, we then have

$$(1 - \xi_{i,i}L)\text{Cov}(\epsilon_t^{2i}, \epsilon_{t-k}^{2m}) = \sum_{j=1}^{i-1} \xi_{i,j} \text{Cov}(\epsilon_{t-1}^{2j}, \epsilon_{t-k}^{2m}), \quad k > 1 \qquad \text{(A.3)}$$

Note the right hand side of (A.3) is equal to equal to zero when $i = 1$. Applying to this equality the lag polynomial $1 - \xi_{i-1,i-1}L$ we get

$$\begin{aligned}(1 - \xi_{i,i}L)(1 - \xi_{i-1,i-1}L)\text{Cov}(\epsilon_t^{2i}, \epsilon_{t-k}^{2m}) &= \xi_{i,i-1}(1 - \xi_{i-1,i-1}L)\text{Cov}(\epsilon_{t-1}^{2(i-1)}, \epsilon_{t-k}^{2m}) \\ &\quad + \sum_{j=1}^{i-2} \xi_{i,j}(1 - \xi_{i-1,i-1}L)\text{Cov}(\epsilon_{t-1}^{2j}, \epsilon_{t-k}^{2m}).\end{aligned}$$

Since by stationarity $\text{Cov}(\epsilon_{t-1}^{2(i-1)}, \epsilon_{t-k}^{2m}) = \text{Cov}(\epsilon_t^{2(i-1)}, \epsilon_{t-k+1}^{2m})$, we can use (A.3) to obtain, for $k > 2$,

$$\begin{aligned}&(1 - \xi_{i,i}L)(1 - \xi_{i-1,i-1}L)\text{Cov}(\epsilon_t^{2i}, \epsilon_{t-k}^{2m}) \\ = \;& \xi_{i,i-1} \sum_{j=1}^{i-2} \xi_{i-1,j} \text{Cov}(\epsilon_{t-2}^{2j}, \epsilon_{t-k}^{2m}) + \sum_{j=1}^{i-2} \xi_{i,j}(1 - \xi_{i-1,i-1}L)\text{Cov}(\epsilon_{t-1}^{2j}, \epsilon_{t-k}^{2m}) \\ = \;& \sum_{j=1}^{i-2} (\xi_{i,i-1}\xi_{i-1,j} - \xi_{i-1,i-1}\xi_{i,j})\text{Cov}(\epsilon_{t-2}^{2j}, \epsilon_{t-k}^{2m}) + \sum_{j=1}^{i-2} \xi_{i,j} \text{Cov}(\epsilon_{t-1}^{2j}, \epsilon_{t-k}^{2m}).\end{aligned}$$

The right hand side of previous equality is zero when $i = 2$. Applying iteratively equation (A.3) we obtain, for $k > i$,

$$\prod_{j=1}^{i}(1-\xi_{j,j}L)\mathrm{Cov}(\epsilon_t^{2i}, \epsilon_{t-k}^{2m}) = 0.$$

In particular

$$\prod_{j=1}^{m}(1-\xi_{j,j}L)\mathrm{Cov}(\epsilon_t^{2m}, \epsilon_{t-k}^{2m}) = 0, \qquad k > m. \tag{A.4}$$

In view of the standard result characterizing the existence of an ARMA representation through the autocovariance function (see *e.g.* Brockwell and Davis (1991), Proposition 3.2.1 and Remark p. 90), we conclude that (ϵ_t^{2m}) admits an ARMA(m, m) representation as claimed. □

Derivation of the ARMA(m, m) representation of Theorem 12.1. By (A.4) the AR polynomial of the ARMA representation is $\prod_{j=1}^{m}(1-\xi_{j,j}L)$. Notice that, by (A.1),

$$(1-\xi_{i,i})E(\epsilon_t^{2i}) = \sum_{j=0}^{i-1}\xi_{i,j}E(\epsilon_{t-1}^{2j}),$$

which proves that $\xi_{i,i} < 1$ for $i = 1, \ldots, m$. The condition $\xi_{i,i} < 1$ can be shown to be necessary and sufficient for the existence of a strictly stationary nonanticipative solution of Model (12.4) with $E\epsilon_t^{2i} < \infty$.

The MA part can be derived by computing the first $m+1$ autocovariances of ϵ_t^{2m}. The moments of ϵ_t^2 up to the order $2m$ are obtained recursively from (A.1). Hence the variance of ϵ_t^{2m}. Next, the first-order autocovariance can be obtained from

$$E(\epsilon_t^{2i}\epsilon_{t-1}^{2m}) = \sum_{j=0}^{i}\binom{i}{j}\omega^{i-j}E\{(\alpha\eta_{t-1}^2+\beta)^j\eta_{t-1}^{2m}\}\frac{\mu_{2i}}{\mu_{2(m+j)}}E\epsilon_{t-1}^{2(m+j)},$$

implying, for $i = 1, \ldots, m$

$$\begin{aligned}\mathrm{Cov}(\epsilon_t^{2i}, \epsilon_{t-1}^{2m}) = \sum_{j=1}^{i}\binom{i}{j}\omega^{i-j}\mu_{2i}\Big\{&E\{(\alpha\eta_{t-1}^2+\beta)^j\eta_{t-1}^{2m}\}\frac{1}{\mu_{2(m+j)}}E\epsilon_t^{2(m+j)}\\ &-E(\alpha\eta_{t-1}^2+\beta)^j\frac{1}{\mu_{2j}}E\epsilon_t^{2j}E\epsilon_t^{2m}\Big\}.\end{aligned}$$

The other autocovariances follow from (A.2).

For instance if the ARMA(1,1) representation (12.5) for ϵ_t^2 was not known, it could be obtained as follows. First note that $\xi_{1,1} = \alpha + \beta$, which gives the AR coefficient. To derive the MA coefficient we derive the first-order autocorrelation of ϵ_t^2. From

$$E(\epsilon_t^2) = \frac{\omega}{1-(\alpha+\beta)}, \quad E(\epsilon_t^4) = \frac{\omega^2\mu_4(1+\alpha+\beta)}{\{1-(\alpha+\beta)\}\{1-(\alpha+\beta)^2-(\mu_4-1)\alpha^2\}},$$

we get

$$\begin{aligned}\gamma(0) &:= \mathrm{Var}(\epsilon_t^2) = \frac{\omega^2(\mu_4-1)(1-2\alpha\beta-\beta^2)}{\{1-(\alpha+\beta)\}^2\{1-(\alpha+\beta)^2-(\mu_4-1)\alpha^2\}}\\ \gamma(1) &:= \mathrm{Cov}(\epsilon_t^2,\epsilon_{t-1}^2) = \frac{\omega^2\alpha(\mu_4-1)(1-(\alpha+\beta)\beta)}{\{1-(\alpha+\beta)\}^2\{1-(\alpha+\beta)^2-(\mu_4-1)\alpha^2\}}\\ \rho(1) &:= \frac{\gamma(1)}{\gamma(0)} = \frac{\alpha(1-(\alpha+\beta)\beta)}{1-2\alpha\beta-\beta^2}.\end{aligned}$$

For an ARMA(1,1) with AR coefficient ϕ and MA coefficient θ, the first-order autocorrelation is

$$\rho(1) = \frac{(\phi-\theta)(1-\theta\phi)}{1-2\theta\phi+\theta^2}.$$

Given that $\phi = \alpha+\beta$ we obtain $\theta = \beta$ or, if $\beta \neq 0$, $\theta = 1/\beta$. Since $0 \le \beta < 1$, the canonical ARMA representation is therefore given by (12.5). When $\beta = 0$ we obtain an AR(1) model.

Proof of Theorem 12.2. For $h \ge 0$, let $\gamma^*(h) = \gamma^*(-h) = \frac{1}{n}\sum_{t=1}^n X_t X_{t+h}$. We have

$$n\mathrm{Cov}\{\hat\gamma(h),\hat\gamma(k)\} - n\mathrm{Cov}\{\gamma^*(h),\gamma^*(k)\} = d_1+d_2+d_3,$$

where $d_1{=}n\mathrm{Cov}\{\hat\gamma(h)-\gamma^*(h),\gamma^*(k)\}$, $d_2{=}n\mathrm{Cov}\{\gamma^*(h),\hat\gamma(k)-\gamma^*(k)\}$ and $d_3{=}n\mathrm{Cov}$ $\{\hat\gamma(h)-\gamma^*(h),\hat\gamma(k)-\gamma^*(k)\}$. By the fourth-order stationarity,

$$|d_1| \quad = \quad n\left|\mathrm{Cov}\left(\frac{1}{n}\sum_{t=n-h+1}^{n} X_tX_{t+h}, \frac{1}{n}\sum_{s=1}^{n} X_sX_{s+k}\right)\right| \le \frac{h}{n}\sum_{\ell=-\infty}^{+\infty}|\sigma_{h,k}(\ell)| \to 0.$$

Similarly, it can be shown that d_2 and d_3 tend to zero as $n\to\infty$.

This entails that we can replace $\{\hat\gamma(h),\hat\gamma(k)\}$ by $\{\gamma^*(h),\gamma^*(k)\}$ to show the theorem. By stationarity, we have

$$\begin{aligned}n\mathrm{Cov}\{\gamma^*(h),\gamma^*(k)\} &= \frac{1}{n}\sum_{t,s=1}^{n}\mathrm{Cov}\left(X_tX_{t+h},X_sX_{s+k}\right)\\ &= \frac{1}{n}\sum_{\ell=-n+1}^{n-1}(n-|\ell|)\mathrm{Cov}\left(X_1X_{1+h},X_{1+\ell}X_{1+\ell+k}\right) \to \sum_{\ell=-\infty}^{+\infty}\sigma_{h,k}(\ell)\end{aligned}$$

using the dominated convergence theorem. □

Proof of Theorem 12.3. The result is obvious for $h=1$. For $h>1$, we have

$$\hat r(h) = \frac{\hat\rho(h) - \sum_{i=1}^{h-1}\hat\rho(h-i)\hat a_{h-1,i}}{1-\sum_{i=1}^{h-1}\hat\rho(i)\hat a_{h-1,i}},$$

where $(\hat a_{h-1,1},\ldots,\hat a_{h-1,h-1})'$ is the vector of the estimated coefficients in the regression of X_t on X_{t-1}, $\ldots$, X_{t-h+1}, $(t=h,\ldots,n)$. Using the ergodic theorem (stationarity and **A1** imply ergodicity), it can be shown that, $\hat\rho(1)\to 0,\ldots,\hat\rho(h)\to 0$ and $(\hat a_{h-1,1},\ldots,\hat a_{h-1,h-1})'\to 0$ almost surely. From Romano and Thombs (1996), we know that $\sqrt{n}\hat\rho(h) = O_P(1)$. Similarly, it can be shown that $\sqrt{n}\left(\hat a_{h-1,1},\ldots,\hat a_{h-1,h-1}\right)$ $= O_P(1)$. Therefore $\hat\rho(k)\hat a_{h-1,i} = O_P(n^{-1})$ for $i=1,\ldots,h-1$ and $k=1,\ldots,h$. Thus

$$n\left\{\hat\rho(h)-\hat r(h)\right\} = \frac{n\sum_{i=1}^{h-1}\hat a_{h-1,i}\left\{\hat\rho(h-i)-\hat\rho(i)\hat\rho(h)\right\}}{1-\sum_{i=1}^{h-1}\hat\rho(i)\hat a_{h-1,i}} = O_p(1). \quad \square$$

Proof of Theorem 12.4. It can be shown that $\Sigma_{\hat{\rho}_{1:m}}$ is non-singular. So the proof comes from (12.7), Remark 12.5, and a standard result (see *e.g.* Brockwell and Davis, 1991, Problem 6.14). □

Proof of Theorem 12.5. We only give the scheme of the proof. A detailed proof is given in an unpublished document (a preliminary version of Francq and Zakoïan, 1998a) which is available on request. It can be shown that the identifiability assumptions, in **A3** − **A6** entail that for all $\theta \in \Theta$ and all $t \in \mathbb{Z}$,

$$\epsilon_t(\theta) = \epsilon_t \ a.s. \quad \Rightarrow \quad \theta = \theta_0. \tag{A.5}$$

It can also be shown that the assumptions on the roots of the AR and MA polynomials, in **A3**, entail that the initial values $e_0(\theta), \ldots, e_{-q+1}(\theta), X_0, \ldots, X_{-p+1}$ are asymptotically negligible. More precisely, we show that

$$\sup_{\theta\in\Theta^*} |c_i(\theta)| = O(\rho^i) \quad \text{for some } \rho \in [0,1[, \quad \lim_{t\to\infty} \sup_{\theta\in\Theta^*} |\epsilon_t(\theta) - e_t(\theta)| = 0,$$

$$\lim_{n\to\infty} \sup_{\theta\in\Theta^*} |Q_n(\theta) - O_n(\theta)| = 0 \quad \text{where} \quad O_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \epsilon_t^2(\theta). \tag{A.6}$$

Since $\epsilon_t - \epsilon_t(\theta) = \sum_{i=1}^{\infty} \{c_i(\theta_0) - c_i(\theta)\} X_{t-i}$ belongs to the linear past of X_t, it is uncorrelated with the linear innovation ϵ_t. Hence the limit criterion $Q_\infty(\theta) := E_{\theta_0}\epsilon_t^2(\theta)$ satisfies $Q_\infty(\theta) = E_{\theta_0}\{\epsilon_t(\theta) - \epsilon_t + \epsilon_t\}^2 = E_{\theta_0}\{\epsilon_t(\theta) - \epsilon_t\}^2 + E_{\theta_0}\epsilon_t^2 + 2\mathrm{Cov}\{\epsilon_t(\theta) - \epsilon_t, \epsilon_t\} = E_{\theta_0}\{\epsilon_t(\theta) - \epsilon_t\}^2 + \sigma^2 \geq \sigma^2$, with equality if and only if $\epsilon_t(\theta) = \epsilon_t$ a.s. In view of (A.5), the last equality holds if and only if $\theta = \theta_0$. Thus, we have shown that $Q_\infty(\theta)$ is minimized at θ_0:

$$\sigma^2 = Q_\infty(\theta_0) < Q_\infty(\theta), \quad \forall\theta \neq \theta_0. \tag{A.7}$$

Let $V_m(\theta^*)$ be the sphere with center θ^* and radius $1/m$. It is clear that $S_m(t) := \inf_{\theta\in V_m(\theta^*)\cap\Theta} \epsilon_t^2(\theta)$ is measurable and integrable. The process $\{S_m(t)\}_t$ is stationary and ergodic. The ergodic theorem shows that, almost surely,

$$\inf_{\theta\in V_m(\theta^*)\cap\Theta} O_n(\theta) = \inf_{\theta\in V_m(\theta^*)\cap\Theta} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2(\theta) \geq \frac{1}{n}\sum_{t=1}^{n} S_m(t) \to E_{\theta_0} S_m(t),$$

as $n \to \infty$. Since $\epsilon_t^2(\theta)$ is continuous in θ, $S_m(t)$ increases to $\epsilon_t^2(\theta^*)$ as m increases to $+\infty$. By Beppo-Levi's theorem we obtain

$$\lim_{m\to\infty} E_{\theta_0} S_m(t) = E_{\theta_0}\epsilon_t^2(\theta^*) = Q_\infty(\theta^*).$$

From (A.7), we then have

$$\liminf_{m\to\infty}\ \liminf_{n\to\infty}\ \inf_{\theta\in V_m(\theta^*)} O_n(\theta) \geq Q_\infty(\theta^*) > \sigma^2 \quad \forall\theta^* \in \Theta,\ \theta^* \neq \theta_0.$$

Thus we have shown that, for all $\theta^* \in \Theta$, $\theta^* \neq \theta_0$, there exists a neighborhood $V(\theta^*)$ of θ^* such that $V(\theta^*) \subset \Theta$ and

$$\liminf_{n\to\infty}\ \inf_{\theta\in V(\theta^*)} O_n(\theta) > \sigma^2, \quad a.s. \tag{A.8}$$

In view of (A.6), (A.8) and the inequality

$$\inf_{\theta\in\theta\in\Theta^*} Q_n(\theta) \geq \inf_{\theta\in\theta\in\Theta^*} O_n(\theta) - \sup_{\theta\in\Theta^*} |O_n(\theta) - Q_n(\theta)|$$

we obtain that, for all $\theta^* \in \Theta^*$, $\theta^* \neq \theta_0$, there exists a neighborhood $V(\theta^*)$ of θ^* such that $V(\theta^*) \subset \Theta$ and

$$\liminf_{n\to\infty} \inf_{\theta\in V(\theta^*)} Q_n(\theta) > \sigma^2, \quad a.s. \tag{A.9}$$

We conclude by a standard compactness argument. Let $V(\theta_0)$ be a neighborhood of θ_0. The compact set Θ^* is covered by $V(\theta_0)$ and the union of the open sets $V(\theta^*)$, $\theta^* \in \Theta^* - V(\theta_0)$, where the $V(\theta^*)$'s satisfy (A.9). Therefore Θ^* is covered by a finite number of these open sets: there exist $\theta_1, ..., \theta_k$ such that $\bigcup_{i=0}^k V(\theta_i) \subset \Theta^*$. Inequality (A.9) shows that, almost surely,

$$\inf_{\theta\in\Theta^*} Q_n(\theta) = \min_{i=0,1,\dots,k} \inf_{\theta\in V(\theta_i)\cap\Theta^*} Q_n(\theta) = \inf_{\theta\in V(\theta_0)\cap\Theta^*} Q_n(\theta),$$

for sufficiently large n. Thus, almost surely, $\hat{\theta}_n$ belongs to $V(\theta_0)$ for sufficiently large n. Since $V(\theta_0)$ can be chosen arbitrarily small, the proof is complete. □

Proof of Theorem 12.6. Under Assumption **A1**, the proof is given in Francq and Zakoïan (1998a). Under Assumption **A1'**, the scheme of the proof is the same. It relies on a Taylor expansion of the criterion $Q_n(\theta)$ around θ_0, and on showing that

$$\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_0) \overset{d}{\rightsquigarrow} \mathcal{N}(0, 4I). \tag{A.10}$$

To show (A.10), note that $\partial\epsilon_t(\theta_0)/\partial\theta = \sum_{i=1}^{\infty} \epsilon_{t-i}\lambda_i$ for some sequence of vectors (λ_i) (see Remark 12.7 below for an explicit form of these vectors). Thus, we can write

$$\begin{aligned} n^{1/2}\frac{\partial}{\partial\theta}Q_n(\theta_0) &= n^{-1/2}\sum_{t=2}^{n} Y_{t,r} + n^{-1/2}\sum_{t=2}^{n} Z_{t,r} + o_P(1), \\ Y_{t,r} &= 2\epsilon_t \sum_{i=1}^{r} \lambda_i\epsilon_{t-i}, \quad Z_{t,r} = 2\epsilon_t \sum_{i=r+1}^{\infty} \lambda_i\epsilon_{t-i}. \end{aligned}$$

Under **A1'**, the process $Y_r := (Y_{t,r})_t$ is strongly mixing, since $\alpha_{Y_r}(h) \leq \alpha_\epsilon(h - r - 1)$. From the central limit theorem for mixing processes (see Herrndorf, 1984), we show that $n^{-1/2}\sum_{t=2}^{n} Y_{t,r} \overset{d}{\rightsquigarrow} \mathcal{N}(0, 4I_r)$ where $I_r \to I$ as $r \to \infty$. Using the arguments given in Francq and Zakoïan (1998a, Lemma 4), it can be shown that $\lim_{r\to\infty} \sup_n \mathrm{Var}\left\{n^{-1/2}\sum_{t=2}^{n} Z_{t,r}\right\} = 0$. So the proof comes from a standard result (see *e.g.* Billingsley, 1995, Theorem 12.5). □

References

Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.

Béguin, J.M., Gouriéroux, C., and Monfort, A. (1980). Identification of a mixed autoregressive-moving average process: The corner method. In: Anderson, O.D. (ed.), *Time Series*, pages 423–436, North-Holland, Amsterdam.

Berk, K.N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics*, 2:489–502.

Berlinet, A. (1984). Estimating the degrees of an ARMA model. *Compstat Lectures*, 3:61–94.

Berlinet, A., and Francq, C. (1997). On Bartlett's Formula for nonlinear processes. *Journal of Time Series Analysis*, 18:535–55.

Berlinet, A., and Francq, C. (1999). Estimation des covariances entre autocovariances empiriques de processus multivariés non linéaires. *La revue Canadienne de Statistique*, 27:1–22.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York.

Box, G.E.P., and Pierce, D.A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *JASA, Journal of the American Statistical Association*, 65:1509–1526.

Brockwell, P.J., and Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer-Verlag.

Broze, L., Francq, C., and Zakoïan, J-M. (2002). Efficient use of higher-lag autocorrelations for estimating autoregressive processes, *Journal of Time Series Analysis*, 23:287–312.

Carrasco, M., and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, 18:17–39.

Dickey, D.A., and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *JASA, Journal of the American Statistical Association*, 74:427–431.

Dunsmuir, W., and Hannan, E.J. (1976). Vector linear time series models. *Advances in Applied Probability*, 8:339–364.

Francq, C., and Zakoïan, J-M. (1998a). Estimating linear representations of nonlinear processes. *Journal of Statistical Planning and Inference*, 68:145–165.

Francq, C., and Zakoïan, J.-M. (1998b). Estimating the order of weak ARMA models. *Prague Stochastic'98 Proceedings*, 1:165–168.

Francq, C., and Zakoïan, J.-M. (2000). Covariance matrix estimation for estimators of mixing Wold's ARMA. *Journal of Statistical Planning and Inference*, 83:369–394.

Francq, C., and Zakoïan, J.-M. (2003). The L^2-structures of standard and switching-regime GARCH models. *Working document*.

Francq, C., Roy R., and Zakoïan, J.-M. (2003). Goodness-of-fit tests for ARMA models with uncorrelated errors. *Working document*.

Glasbey, C.A. (1982). Generalization of partial autocorrelation useful in identifying ARMA models. *Technometrics*, 24:223–228.

Hannan, E.J. (1973). The asymptotic theory of linear time series models. *Journal of Applied Probability*, 10:130–145.

Hansen, B.E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica*, 60:967–972.

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. *Annals of Probability*, 12:141–153.

Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426.

Kokoszka, P.S., and Taqqu, M.S. (1996). Parameter estimation for infinite variance fractional ARIMA. *The Annals of Statistics*, 24:1880–1913.

Ljung, G.M., and Box, G.E.P. (1978). On the Measure of lack of fit in time series models. *Biometrika*, 65:297–303.

Lobato, I.N. (2001). Testing that a dependent process is uncorrelated. *JASA, Journal of the American Statistical Association*, 96:1066–1076.

Mann, H.B., and Wald, A. (1943). On the statistical treatment of linear stochastic differential equations. *Econometrica*, 11:173–200.

Mélard, G., and Roy, R. (1987). On confidence intervals and tests for autocorrelations. *Computational Statistics and Data Analysis*, 5:31–44.

Mélard, G., Paesmans, M., and Roy, R. (1991). Consistent estimation of the asymptotic covariance structure of multivariate serial correlations. *Journal of Time Series Analysis*, 12:351–361.

Mikosch, T., Gadrich, T., Klüppelberg, C., and Adler, R.J. (1995). Parameter estimation for ARMA models with infinite variance innovations. *The Annals of Statistics*, 23:305–326.

Pham, D.T. (1986). The mixing property of bilinear and generalized random coefficient autoregressive models. *Stochastic Processes and their Applications*, 23:291–300.

Rissanen, J., and Caines, P.E. (1979). The strong consistency of maximum likelihood estimators for ARMA processes. *The Annals of Statistics*, 7:297–315.

Romano, J.L., and Thombs, L.A. (1996). Inference for autocorrelations under weak assumptions. *JASA, Journal of the American Statistical Association*, 91:590–600.

Roy, R. (1989). Asymptotic covariance structure of serial correlations in multivariate time series. *Biometrika*, 76:824-827.

Tiao, G.C., and Tsay, R.S. (1983). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *The Annals of Statistics*, 11:856–871.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach.* Clarendon Press Oxford.

Walker, A.M. (1964). Asymptotic properties of least squares estimates of the parameters of the spectrum of a stationary non-deterministic time series. *Journal of the Australian Mathematical Society*, 4:363–384.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series.* Almqvist and Wiksell, Stocholm.