CRM-3371

On strong consistency and asymptotic normality of one-step Gauss-Newton estimators in ARMA time series models *

Pierre Duchesne*, Pierre Lafaye de Micheaux†, Joseph François Tagne Tatsinkou*
*Département de mathématiques et de statistique
Université de Montréal, Montréal, Canada
†School of Mathematics and Statistics
University of New South Wales, Sydney, Australia

Abstract

We provide a proof of the strong consistency and we study the asymptotic normality of one-step Gauss-Newton estimators for causal and invertible autoregressive moving-average time series models. In a small simulation study, the empirical properties of several estimators are illustrated in an order-one moving-average model. We compared empirically the following methods: the estimators based on the moment method, based on the innovations algorithm procedure, and using the one-step Gauss-Newton estimators relying on these preliminary estimators (using mean-corrected data, and also using estimators of the mean and the moving-average parameter as starting points). These estimators are compared empirically to the maximum likelihood estimator with respect to exact biases and mean squared errors.

Keywords: Asymptotic normality; innovations algorithm; method of moments; one-step Gauss-Newton estimators; strong consistency.

1. Introduction

Let $\{Y_t, t \in \mathbb{Z}\}$ be a stationary, invertible and identifiable autoregressive moving-average (ARMA) stochastic process satisfying the difference equation:

$$\Phi_p(\boldsymbol{\phi}, B)(Y_t - \mu) = \Theta_q(\boldsymbol{\theta}, B)\epsilon_t, \tag{1}$$

where $\Phi_p(\boldsymbol{\phi}, B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\Theta_q(\boldsymbol{\theta}, B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are the autoregressive and moving-average polynomials, respectively, with B being the backshift

*Corresponding author: Pierre Duchesne, Université de Montréal; Département de mathématiques et de statistique; C.P. 6128 Succursale Centre-Ville; Montréal, Québec H3C 3J7; Canada. tel: (514) 343-7267; fax: (514) 343-5700; e-mail: duchesne@dms.umontreal.ca.

operator satisfying $BY_t = Y_{t-1}$. The error term $\{\epsilon_t, t \in \mathbb{Z}\}$ is composed of independent and identically distributed (iid) random variables with $E(\epsilon_t) = 0$ and a finite variance $\text{var}(\epsilon_t) = \sigma^2 > 0$. In this article, the autoregressive and moving-average orders are supposed known. Identification techniques for p and q are discussed in Brockwell and Davis (1991, Sect. 9.2). The stochastic process $\{Y_t, t \in \mathbb{Z}\}$ satisfying (1) is said ARMA(p,q) with mean $E(Y_t) = \mu$. The mean parameter μ is assumed finite.

The model parameters are collected in the vector

$$\boldsymbol{\nu} = (\boldsymbol{\beta}^\top, \sigma^2)^\top,$$

where $\boldsymbol{\beta} = (\boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top, \mu)^\top$, with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$. Note that μ is included in $\boldsymbol{\beta}$. It will be useful to have a notation for the model parameters without the mean, denoted $\boldsymbol{\beta}_{(\mu)} = (\boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top)^\top$. It is assumed that $\boldsymbol{\beta}_{(\mu)}$ and μ are not functionally related; thus $\boldsymbol{\beta}_{(\mu)}$ is not function of μ, and μ is not function of $\boldsymbol{\beta}_{(\mu)}$. Estimators of $\boldsymbol{\nu}$ with desirable properties (such as weak consistency and asymptotic normality) can be obtained by maximizing the likelihood of a stationary Gaussian time series with respect to $\boldsymbol{\nu}$. Since normality of the errors is not always assumed here, these estimators are called maximum Gaussian likelihood estimators in Brockwell and Davis (1991). The properties of these estimators go back to Hannan (1973) via the asymptotic properties of a Whittle's estimator. The asymptotic properties are also studied in Brockwell and Davis (1991) and Fuller (1996). Direct proofs of strong consistency and asymptotic normality are given in Yao and Brockwell (2006).

In general, the likelihood function of a finite realization of an ARMA model is non-linear. It relies on the determinant of the covariance matrix of the process and also on a quadratic form involving the inverse of the covariance matrix. Typically, in order to compute the maximum likelihood estimators, algorithms require initial values. Brockwell and Davis (2002) discuss several techniques, notably the method of moments (or Yule-Walker estimation when $q = 0$), Burg's algorithm, estimators obtained from the Brockwell-Davis innovations algorithm, and the Hannan-Rissanen algorithm. Except for the innovations algorithm, all the other estimation techniques are particularly useful when $q = 0$. Brockwell and Davis (1991, Section 8.3) discuss how the innovations algorithm can be used to estimate MA(q) models. Note that when $q > 0$, the method of moments and the innovations algorithm give inefficient estimators (Brockwell and Davis, 1991, Section 8.5). Algorithms to compute efficiently the likelihood function are discussed in Luceño (1993).

Another technique when preliminary estimators are available is one-step Gauss-Newton (1SGN) estimation. That technique has a long history in statistics, see for example Ferguson (1996, pp. 134-138) or van der Vaart (1998, Section 5.7). In ARMA modelling, it is discussed in MA(1) models in Fuller (1996, pp. 422-425) and, more generally, for ARMA models in Brockwell and Davis (1991, Section 8.11). Under weak conditions of the preliminary estimators (which are given in Section 2), 1SGN estimators are relatively simple to

compute and they attain asymptotic efficiency when the normality assumption holds true, even if the preliminary estimator is inefficient. Here, we give a detailed proof of the strong consistency of 1SGN estimators, which was not available in the literature, to the best of our knowledge. Asymptotic normality is also studied. In particular, the first order expansion satisfied by the maximum likelihood estimator is justified under precise conditions for 1SGN estimators.

The rest of this paper is organized as follows. Some technical preliminaries are given in Section 2. Section 3 establishes the strong consistency of the 1SGN estimators, and Section 4 studies asymptotic normality. A small simulation study is presented in Section 5. The properties of the 1SGN estimators are illustrated in MA(1) models. The following methods are compared to estimate μ and θ: the estimators calculated by the method of moments and by the innovations algorithm; the 1SGN estimators based on mean-corrected data with preliminary estimators calculated using the method of moments and with the innovations algorithm; the 1SGN estimators with preliminary estimators of θ calculated with the method of moments and with the innovations algorithm while μ is estimated by the sample mean; and finally the maximum likelihood estimators of μ and θ. We conclude with a discussion.

2. Preliminaries

We give in this section some technical preliminaries. Throughout the paper, the symbols $\xrightarrow{L}$, $\xrightarrow{P}$, $\xrightarrow{a.s.}$ stand for convergence in distribution, in probability, and almost sure convergence, respectively. Let X_n be a random variable. The notation $X_n = o_P(1)$ denotes $X_n \xrightarrow{P} 0$ and $X_n = O_P(1)$ means bounded in probability. For the random vector $\mathbf{X}_n$, convergence in probability to zero and bounded in probability are noted $\mathbf{X}_n = \mathbf{o}_P(1)$ and $\mathbf{X}_n = \mathbf{O}_P(1)$, respectively. See Serfling (1980) for details. The d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is noted $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The norm of the random variable X satisfies $\|X\|^2 = E(X^2)$. We define the set $\mathcal{I}_n = \{1, \ldots, n\}$ and $\mathbf{I}_n$ corresponds to the identity matrix of order n. We also use K as a generic strictly positive constant, which may differ from place to place, and $\rho \in (0, 1)$.

Let $y_1, \ldots, y_n$ be a finite realization of length n of the stochastic difference equation (1). The vector of true parameters $\boldsymbol{\nu}_0$ in the ARMA(p,q) model (1) is defined as:

$$\boldsymbol{\nu}_0^\top = (\boldsymbol{\beta}_0^\top, \sigma_0^2),$$

where $\boldsymbol{\beta}_0 = (\boldsymbol{\phi}_0^\top, \boldsymbol{\theta}_0^\top, \mu_0)^\top$, $\boldsymbol{\phi}_0 = (\phi_{01}, \ldots, \phi_{0p})^\top$ and $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0q})^\top$. We assume that $\boldsymbol{\nu}_0$ satisfies Assumption A.

Assumption A. *Introduce the following set:*

$$\begin{aligned}\mathcal{C}_\delta \;=\; \{&\boldsymbol{\beta}_{(\mu)} \in \mathbb{R}^{p+q} \mid \textit{The polynomials } \Phi_p(\boldsymbol{\phi}, z) \textit{ and } \Theta_q(\boldsymbol{\theta}, z) \textit{ have no common roots} \\ &\textit{and all their zeros are outside the unit disk with moduli} \geq 1 + \delta\}.\end{aligned}$$

It is supposed that the vector $\boldsymbol{\beta}_{0(\mu)}$ belongs to the parameter space $\mathcal{C}_\delta$ for some $\delta > 0$.

The introduction of δ in Assumption A is similar to an hypothesis made in Francq and Zakoïan (1998, 2000). See also Yao and Brockwell (2006, p. 859). That technical assumption assures that $\boldsymbol{\beta}_{(\mu)}$ lies in a compact set. The hypothesis stating that the polynomials $\Phi_p(\boldsymbol{\phi}, z)$ and $\Theta_q(\boldsymbol{\theta}, z)$ have no common roots is also called a non redundancy condition in McLeod (1993). For all μ and $\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta$, let $\{\epsilon_t(\boldsymbol{\beta}), t \in \mathbb{Z}\}$ be the second order stationary process defined as the solution of:

$$\epsilon_t(\boldsymbol{\beta}) = \Phi_p(\boldsymbol{\phi}, B)(Y_t - \mu) - \sum_{j=1}^{q} \theta_j \epsilon_{t-j}(\boldsymbol{\beta}). \tag{2}$$

Note that $\epsilon_t(\boldsymbol{\beta}_0) = \epsilon_t$, *a.s.*, $\forall t \in \mathbb{Z}$. There exist weights $\{\pi_j(\boldsymbol{\beta}_{(\mu)}), j \in \mathbb{N}\}$ such that $\sum_{j=0}^{\infty} |\pi_j(\boldsymbol{\beta}_{(\mu)})| < \infty$, and $\{\epsilon_t(\boldsymbol{\beta})\}$ can be written as:

$$\epsilon_t(\boldsymbol{\beta}) = (Y_t - \mu) + \sum_{j=1}^{\infty} \pi_j(\boldsymbol{\beta}_{(\mu)})(Y_{t-j} - \mu). \tag{3}$$

There also exist absolutely summable weights $\{\psi_j(\boldsymbol{\beta}_{(\mu)}), j \in \mathbb{N}\}$ satisfying the condition $\sum_{j=0}^{\infty} |\psi_j(\boldsymbol{\beta}_{(\mu)})| < \infty$, such that $\{Y_t\}$ can be written as the linear process:

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j(\boldsymbol{\beta}_{(\mu)}) \epsilon_{t-j}(\boldsymbol{\beta}). \tag{4}$$

The representations (3) and (4) are said invertible and causal, respectively. Note that $\epsilon_t(\cdot)$ is a continuous function of $\boldsymbol{\beta}$ but the weights $\{\pi_j(\boldsymbol{\beta}_{(\mu)})\}$ and $\{\psi_j(\boldsymbol{\beta}_{(\mu)})\}$ are function of $\boldsymbol{\beta}_{(\mu)}$ only. For any finite μ and $\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta$, $\{\epsilon_t(\boldsymbol{\beta}), t \in \mathbb{Z}\}$ defines an ergodic sequence and $E\{\epsilon_t^2(\boldsymbol{\beta})\} < \infty$. See also Francq and Zakoïan (1998, 2000) for a similar framework.

For any μ and $\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta$ with $t \in \mathcal{I}_n$, the residuals $e_t(\boldsymbol{\beta})$ are calculated recursively according to the relations:

$$e_t(\boldsymbol{\beta}) = \begin{cases} Y_1 - \mu, & \text{if } t = 1, \\ (Y_2 - \mu) - \phi_1(Y_1 - \mu) - \theta_1 e_1(\boldsymbol{\beta}), & \text{if } t = 2, \\ \vdots & \\ (Y_n - \mu) - \sum_{j=1}^{p} \phi_j(Y_{n-j} - \mu) - \sum_{j=1}^{q} \theta_j e_{n-j}(\boldsymbol{\beta}), & \text{if } t = n. \end{cases}$$

The residuals are well-defined for all integers up to n by posing the initial values $e_t(\boldsymbol{\beta}) = 0$ for all $t \leq n$. Setting $Y_t = \mu$, $t \leq 0$, the process $\{Y_t\}$ can be written as:

$$\Phi_p(\boldsymbol{\phi}, B)(Y_t - \mu) = \Theta_q(\boldsymbol{\theta}, B) e_t(\boldsymbol{\beta}),$$

for all $t \leq n$. See Brockwell and Davis (1991, p. 195). It follows that for $t \leq n$:

$$e_t(\boldsymbol{\beta}) = \Theta_q^{-1}(\boldsymbol{\theta}, B)\Phi_p(\boldsymbol{\phi}, B)(Y_t - \mu) = (Y_t - \mu) + \sum_{j=1}^{t-1} \pi_j(\boldsymbol{\beta}_{(\mu)})(Y_{t-j} - \mu). \tag{5}$$

Intuitively, the random variables $e_t(\boldsymbol{\beta}_0)$, $t \leq n$, correspond to idealized residuals, which provide good approximations to the original error variables $\epsilon_t = \epsilon_t(\boldsymbol{\beta}_0)$, $t \leq n$.

One-step Gauss-Newton estimators defined in Brockwell and Davis (1991, p. 265) are investigated to estimate the vector of parameters $\boldsymbol{\nu}_0$. Here, we present their method for a general μ, not only $\mu = 0$. That estimation method relies on a preliminary estimator $\tilde{\boldsymbol{\nu}}_n$, say, of $\boldsymbol{\nu}_0$, which satisfies Assumption B.

Assumption B. *Let $\{Y_t, t \in \mathbb{Z}\}$ be generated by the stochastic difference equation (1). Let $\boldsymbol{\nu}_0$ be the true parameter values in that ARMA(p,q) model. Based on a finite realization of size n, the preliminary estimator $\tilde{\boldsymbol{\nu}}_n = (\tilde{\boldsymbol{\beta}}_n^\top, \tilde{\sigma}_n^2)^\top$ of $\boldsymbol{\nu}_0$ satisfies:*
B(i): the estimator $\tilde{\boldsymbol{\nu}}_n$ is strongly consistent for $\boldsymbol{\nu}_0$;
B(ii): $\tilde{\boldsymbol{\nu}}_n = \boldsymbol{\nu}_0 + \mathbf{o}_P(n^{-1/4})$.

The rate in Assumption B(ii) is weak. Several estimation methods are shown strongly consistent and offer the convergence rate $\mathbf{o}_P(n^{-1/4})$. In fact, any strongly consistent estimator such that $n^{1/2}(\tilde{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_0) = \mathbf{O}_P(1)$ satisfies the specified rate. Examples include least squares estimators and estimators obtained by maximizing the Gaussian likelihood. Based on a preliminary estimator $\tilde{\boldsymbol{\nu}}_n$, define the criterion:

$$S_n(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\nu}}_n) = \sum_{t=1}^{n} \left\{ e_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{d}_t^\top(\tilde{\boldsymbol{\beta}}_n)(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_n) \right\}^2, \tag{6}$$

where $\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) = (d_{t,1}(\tilde{\boldsymbol{\beta}}_n), \ldots, d_{t,p+q+1}(\tilde{\boldsymbol{\beta}}_n))^\top$ corresponds to the $(p+q+1)\times 1$ vector whose components are:

$$d_{t,i}(\tilde{\boldsymbol{\beta}}_n) = -\frac{\partial e_t(\boldsymbol{\beta})}{\partial \beta_i}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}_n}, \ i = 1, \ldots, p+q+1.$$

Note that $\mathbf{d}_t(\boldsymbol{\beta}_0)$ is defined similarly. The one-step Gauss-Newton estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}_{1SGN,n}$, is obtained by minimizing $S_n(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\nu}}_n)$ with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{1SGN,n} = \arg\min_{\boldsymbol{\beta}} S_n(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\nu}}_n).$$

The objective function (6) represents a good approximation of the sum of squares corresponding to the least squares estimators:

$$S_{LS}(\boldsymbol{\beta}) = \sum_{t=1}^{n} \{Y_t - \hat{Y}_t(\boldsymbol{\beta})\}^2 / r_{t-1}(\boldsymbol{\beta}), \tag{7}$$

where $\hat{Y}_t(\boldsymbol{\beta})$ denotes the best linear predictor of Y_t (with minimum mean squared error), function of Y_s, $s \in I_{t-1}$ and r_t is such that $E\{Y_{t+1} - \hat{Y}_t(\boldsymbol{\beta})\}^2 = \sigma^2 r_t$. We refer to Brockwell and Davis (1991) for the detailed expressions.

The criterion (6) is motivated by performing a second-order Taylor expansion of $e_t(\boldsymbol{\beta}_0)$ around $\tilde{\boldsymbol{\beta}}_n$:

$$\begin{aligned} e_t(\boldsymbol{\beta}_0) &= e_t(\tilde{\boldsymbol{\beta}}_n) + \frac{\partial e_t(\tilde{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}^\top}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + \frac{1}{2}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n)^\top \frac{\partial^2 e_t(\boldsymbol{\beta}^*_{tn})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n), \quad (8) \\ &= e_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{d}_t^\top(\tilde{\boldsymbol{\beta}}_n)(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + z_t(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*_{tn}), \quad (9) \end{aligned}$$

where in the first equality $\boldsymbol{\beta}^*_{tn}$ is between $\boldsymbol{\beta}_0$ and $\tilde{\boldsymbol{\beta}}_n$, and the remainder $z_t(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*_{tn})$ satisfies:

$$z_t(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*_{tn}) = \frac{1}{2}\sum_{i=1}^{p+q+1}\sum_{j=1}^{p+q+1} \frac{\partial^2 e_t(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*_{tn}} (\beta_{0i} - \tilde{\beta}_{in})(\beta_{0j} - \tilde{\beta}_{jn}).$$

For notational simplicity we note $z_t \equiv z_t(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*_{tn})$. Combining the equations for $t = 1, \ldots, n$, we can write alternatively the system of equations in a compact form:

$$\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) = \mathbf{D}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) - \mathbf{Z}_n + \mathbf{e}_n(\boldsymbol{\beta}_0), \quad (10)$$

where $\mathbf{e}_n(\boldsymbol{\beta}) = (e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta}))^\top$, $\mathbf{D}_n^\top \equiv \mathbf{D}_n^\top(\tilde{\boldsymbol{\beta}}_n) = (\mathbf{d}_1(\tilde{\boldsymbol{\beta}}_n), \ldots, \mathbf{d}_n(\tilde{\boldsymbol{\beta}}_n))$, $\mathbf{D}_n$ being a $n \times (p+q+1)$ matrix and $\mathbf{Z}_n = (z_1, \ldots, z_n)^\top$. Consider the following regression model in $\boldsymbol{\beta}$:

$$\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) = \mathbf{D}_n(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_n) - \mathbf{Z}_n + \mathbf{e}_n(\boldsymbol{\beta}_0),$$

where $\mathbf{e}_n(\boldsymbol{\beta}_0)$ plays now the role of the error term. In the linear model, the least squares estimators of $\boldsymbol{\beta}$, noted $\boldsymbol{\beta}^\dagger_n$, satisfies:

$$n^{1/2}(\boldsymbol{\beta}^\dagger_n - \tilde{\boldsymbol{\beta}}_n) = n^{1/2}(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) - n^{1/2}(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{Z}_n.$$

Brockwell and Davis (1991, p. 267) showed that $n^{-1}\mathbf{D}_n^\top\mathbf{D}_n = \mathbf{O}_P(1)$ and $\mathbf{D}_n^\top\mathbf{Z}_n = \mathbf{o}_P(n^{1/2})$, implying that $n^{1/2}(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{Z}_n = \mathbf{o}_P(1)$. Neglecting the asymptotically negligible term, the one-step Gauss-Newton estimator represents a one-step improvement of $\tilde{\boldsymbol{\beta}}_n$, defined as:

$$\hat{\boldsymbol{\beta}}_{1SGN,n} = \tilde{\boldsymbol{\beta}}_n + \widehat{\boldsymbol{\Delta}\boldsymbol{\beta}}_n, \quad (11)$$

where $\tilde{\boldsymbol{\beta}}_n$ represents the preliminary estimator of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\Delta}\boldsymbol{\beta}}_n = (\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n)$. The variance σ_0^2 is estimated as:

$$\hat{\sigma}^2_{1SGN,n} = n^{-1}\sum_{t=1}^{n} e_t^2(\tilde{\boldsymbol{\beta}}_{1SGN,n}). \quad (12)$$

Strong consistency and asymptotic normality are studied in the next sections.

3. Strong consistency

Brockwell and Davis (1991, pp. 265-269) studied the asymptotic properties of (11), showing that it is weakly consistent and asymptotically normal. We complete their results by establishing that if the preliminary estimator $\tilde{\boldsymbol{\beta}}_n$ is strongly consistent (but not necessarily efficient), then $\hat{\boldsymbol{\beta}}_{1SGN,n}$ inherits that property, that is $\hat{\boldsymbol{\beta}}_{1SGN,n} \xrightarrow{a.s.} \boldsymbol{\beta}_0$. Note that Assumption B(ii) is not needed to establish the consistency result. Furthermore, $\hat{\sigma}^2_{1SGN,n} \xrightarrow{a.s.} \sigma_0^2$. This is stated more precisely in the next Theorem.

Theorem 1. *Let $\{Y_t, t \in \mathbb{Z}\}$ be a stochastic difference equation generated by (1) at $\boldsymbol{\nu} = \boldsymbol{\nu}_0$, satisfying Assumption A. Let $\tilde{\boldsymbol{\beta}}_n$ be a preliminary estimator which is strongly consistent, that is $\tilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$. Under Assumptions A and B(i), then $\hat{\boldsymbol{\beta}}_{1SGN,n}$ and $\hat{\sigma}^2_{1SGN,n}$ defined by (11) and (12) are strongly consistent:*

$$\hat{\boldsymbol{\nu}}_{1SGN,n} = (\hat{\boldsymbol{\beta}}^\top_{1SGN,n}, \hat{\sigma}^2_{1SGN,n})^\top \xrightarrow{a.s.} \boldsymbol{\nu}_0.$$

Proof. To show that $\hat{\boldsymbol{\beta}}_{1SGN,n} \xrightarrow{a.s.} \boldsymbol{\beta}_0$, it is sufficient to establish the following convergence results:

$$n^{-1}\mathbf{D}_n^\top \mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) \xrightarrow{a.s.} \mathbf{0}, \tag{13}$$

$$n^{-1}\mathbf{D}_n^\top \mathbf{D}_n \xrightarrow{a.s.} \sigma_0^2 \mathbf{V}^{-1}(\boldsymbol{\beta}_0), \tag{14}$$

where $\mathbf{V}(\boldsymbol{\beta}_0) = \sigma_0^2\{E(\mathbf{W}_1\mathbf{W}_1^\top)\}^{-1}$ and $\mathbf{W}_t$ is defined in the statement of Theorem 2. A direct consequence of results (13) and (14) is that $\widehat{\boldsymbol{\Delta}\boldsymbol{\beta}}_n \xrightarrow{a.s.} \mathbf{0}$. Since by hypothesis $\tilde{\boldsymbol{\beta}}_n$ converges almost surely to $\boldsymbol{\beta}_0$, this implies that $\hat{\boldsymbol{\beta}}_{1SGN,n} \xrightarrow{a.s.} \boldsymbol{\beta}_0$. According to Francq and Zakoïan (1998), the weights $\{\pi_i(\boldsymbol{\beta}_{(\mu)})\}$ in the stochastic expansion (3) satisfy:

$$\sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} \left|\pi_i(\boldsymbol{\beta}_{(\mu)})\right| \leq K\rho^i,$$

for certain constants $K > 0$ and $\rho \in (0, 1)$. The following expansions are also valid:

$$\begin{aligned}
\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{\infty} \xi_{ij}(\boldsymbol{\beta}_{(\mu)})(Y_{t-i} - \mu),\ j = 1, \ldots, p+q, \\
\frac{\partial^2 \epsilon_t(\boldsymbol{\beta})}{\partial \beta_{j_1} \partial \beta_{j_2}} &= \sum_{i=1}^{\infty} \xi_{i,j_1 j_2}(\boldsymbol{\beta}_{(\mu)})(Y_{t-i} - \mu),\ j_1, j_2 = 1, \ldots, p+q,
\end{aligned}$$

and the weights $\{\xi_{ij}(\boldsymbol{\beta}_{(\mu)})\}$ and $\{\xi_{i,j_1 j_2}(\boldsymbol{\beta}_{(\mu)})\}$ satisfy:

$$\begin{aligned}
\sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} |\xi_{ij}(\boldsymbol{\beta}_{(\mu)})| &\leq K\rho^i, \\
\sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} |\xi_{i,j_1 j_2}(\boldsymbol{\beta}_{(\mu)})| &\leq K\rho^i,
\end{aligned}$$

where $K > 0$ and $\rho \in (0,1)$ are certain constants. The expansion and the result can be found in Francq and Zakoïan (1998, 2000). For $j = p+q+1$, one easily shows:

$$\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_{p+q+1}} = \frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \mu} = -\left\{1 + \sum_{i=1}^{\infty} \pi_i(\boldsymbol{\beta}_{(\mu)})\right\}.$$

Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be arbitrary vectors. In the sequel, we need to study the stochastic differences $\epsilon_t(\boldsymbol{\beta}_2) - \epsilon_t(\boldsymbol{\beta}_1)$ and $\partial\epsilon_t(\boldsymbol{\beta}_2)/\partial\beta_j - \partial\epsilon_t(\boldsymbol{\beta}_1)/\partial\beta_j$. They are stated in the following lemma.

Lemma 1. *Let $\{\epsilon_t(\boldsymbol{\beta})\}$ be the second order stationary process solution of (2). For any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we have the following stochastic differences:*

$$\begin{aligned}
\epsilon_t(\boldsymbol{\beta}_2) - \epsilon_t(\boldsymbol{\beta}_1) &= (\mu_1 - \mu_2) + \sum_{i=1}^{\infty} \{\pi_i(\boldsymbol{\beta}_{2(\mu)}) - \pi_i(\boldsymbol{\beta}_{1(\mu)})\} Y_{t-i} \\
&+ (\mu_1 - \mu_2) \sum_{i=1}^{\infty} \pi_i(\boldsymbol{\beta}_{1(\mu)}) + \mu_2 \sum_{i=1}^{\infty} \left\{\pi_i(\boldsymbol{\beta}_{1(\mu)}) - \pi_i(\boldsymbol{\beta}_{2(\mu)})\right\}, \qquad (15)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \epsilon_t(\boldsymbol{\beta}_2)}{\partial \beta_j} - \frac{\partial \epsilon_t(\boldsymbol{\beta}_1)}{\partial \beta_j} &= \begin{cases} \sum_{i=1}^{\infty} \{\xi_{i,j}(\boldsymbol{\beta}_{2(\mu)}) - \xi_{i,j}(\boldsymbol{\beta}_{1(\mu)})\} Y_{t-i} + (\mu_1 - \mu_2) \sum_{i=1}^{\infty} \xi_{i,j}(\boldsymbol{\beta}_{1(\mu)}) \\ \quad + \mu_2 \sum_{i=1}^{\infty} \left\{\xi_{ij}(\boldsymbol{\beta}_{1(\mu)}) - \xi_{ij}(\boldsymbol{\beta}_{2(\mu)})\right\}, \text{ if } j = 1, \ldots, p+q, \\ \sum_{i=1}^{\infty} \{\xi_{i,j}(\boldsymbol{\beta}_{1(\mu)}) - \xi_{i,j}(\boldsymbol{\beta}_{2(\mu)})\} Y_{t-i}, \text{ if } j = p+q+1. \end{cases} \qquad (16)
\end{aligned}$$

Proof. The differences are exact. The proofs follow using simple and straightforward algebra. Therefore, they are omitted. □

We now present the following lemma, which is strongly inspired from the results of Francq and Zakoïan (1998, 2000), showing how to include the general case $\mu \neq 0$.

Lemma 2. *Suppose that Assumption A is satisfied. Then*

$$\lim_t \sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} |\epsilon_t(\boldsymbol{\beta}) - e_t(\boldsymbol{\beta})| = 0, a.s. \qquad (17)$$

A direct calculation gives the following equality:

$$\epsilon_t(\boldsymbol{\beta}) - e_t(\boldsymbol{\beta}) = \sum_{i=0}^{\infty} \pi_{i+t}(\boldsymbol{\beta}_{(\mu)})(Y_{-i} - \mu). \qquad (18)$$

Using the Minkowski inequality, we find:

$$\begin{aligned}
|\epsilon_t(\boldsymbol{\beta}) - e_t(\boldsymbol{\beta})| &\leq \sum_{i=0}^{\infty} |\pi_{i+t}(\boldsymbol{\beta}_{(\mu)})||Y_{-i} - \mu| \leq K\rho^t \sum_{i=0}^{\infty} \rho^i (|Y_{-i}| + |\mu|), \\
&= K\rho^t \left\{Z + (1-\rho)^{-1}|\mu|\right\},
\end{aligned}$$

where $Z = \sum_{i=0}^{\infty} \rho^i |Y_{-i}|$. The convergence to zero is fast. In fact,

$$\sum_{t=1}^{\infty} P\left\{\rho^t \left[Z + \frac{|\mu|}{1-\rho}\right] > \epsilon\right\} \leq \frac{1}{\epsilon} \sum_{t=1}^{\infty} \rho^t \left\{E(Z) + \frac{|\mu|}{1-\rho}\right\} < \infty,$$

implying that:

$$\lim_t \sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} |\epsilon_t(\boldsymbol{\beta}) - e_t(\boldsymbol{\beta})| = 0, \ a.s.$$

This generalizes a lemma due to Francq and Zakoïan (1998, 2000) when μ is unknown (but fixed). The following expansions are valid:

$$\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_j} - \frac{\partial e_t(\boldsymbol{\beta})}{\partial \beta_j} = \begin{cases} \sum_{i=0}^{\infty} \{\partial \pi_{i+t}(\partial \boldsymbol{\beta}_{(\mu)})/\partial \beta_j\}(Y_{-i} - \mu), & j = 1, \ldots, p+q, \\ -\sum_{i=t}^{\infty} \pi_i(\boldsymbol{\beta}_{(\mu)}), & j = p+q+1, \end{cases} \tag{19}$$

and similar arguments give:

$$\lim_t \sup_{\boldsymbol{\beta}_{(\mu)} \in \mathcal{C}_\delta} \left| \frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_j} - \frac{\partial e_t(\boldsymbol{\beta})}{\partial \beta_j} \right| = 0, \ a.s., j = 1, \ldots, p+q+1. \tag{20}$$

Remark 1. *The same arguments as in Francq and Zakoïan (1998, 2000) can essentially be used for treating $j = 1, \ldots, p+q$. The case $j = p+q+1$ does not cause new difficulty. Since $|\pi_i(\boldsymbol{\beta}_{(\mu)})| \leq K\rho^i$, for $K > 0$ and $\rho \in (0,1)$, the same kind of arguments allows us to study $\epsilon_t(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})$ or the difference of derivatives $\partial \epsilon_t(\boldsymbol{\beta})/\partial \beta_j - \partial e_t(\boldsymbol{\beta})/\partial \beta_j$.*

First, we study $n^{-1}\mathbf{D}_n^\top \mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n)$ and we show that it converges to zero almost surely. Introduce the $(p+q+1) \times n$ matrix $\mathbf{C}_n$ satisfying $\mathbf{C}_n^\top = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, with

$$\mathbf{c}_t(\boldsymbol{\beta}_0) = (c_{t,1}(\boldsymbol{\beta}_0), \ldots, c_{t,p+q+1}(\boldsymbol{\beta}_0))^\top,$$

the $(p+q+1) \times 1$ vector whose components are given by the derivatives:

$$c_{t,i}(\boldsymbol{\beta}_0) = -\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_i}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}, \ i = 1, \ldots, p+q+1.$$

The following decomposition is used:

$$n^{-1}\mathbf{D}_n^\top \mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) = n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) \ e_t(\tilde{\boldsymbol{\beta}}_n) = \mathbf{S}_{1n} + \mathbf{S}_{2n} + \mathbf{S}_{3n},$$

where

$$\begin{aligned} \mathbf{S}_{1n} &= n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\boldsymbol{\beta}_0) e_t(\boldsymbol{\beta}_0), \\ \mathbf{S}_{2n} &= n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\boldsymbol{\beta}_0) \left\{e_t(\tilde{\boldsymbol{\beta}}_n) - e_t(\boldsymbol{\beta}_0)\right\}, \\ \mathbf{S}_{3n} &= n^{-1} \sum_{t=1}^{n} \left\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{d}_t(\boldsymbol{\beta}_0)\right\} e_t(\tilde{\boldsymbol{\beta}}_n). \end{aligned}$$

First, we consider the term $\mathbf{S}_{1n}$, that we decompose further as $\mathbf{S}_{1n} = \mathbf{S}_{11n} + \mathbf{S}_{12n} + \mathbf{S}_{13n}$, where:

$$\begin{aligned}
\mathbf{S}_{11n} &= n^{-1} \sum_{t=1}^{n} \{\mathbf{d}_t(\boldsymbol{\beta}_0) - \mathbf{c}_t(\boldsymbol{\beta}_0)\} e_t(\boldsymbol{\beta}_0), \\
\mathbf{S}_{12n} &= n^{-1} \sum_{t=1}^{n} \mathbf{c}_t(\boldsymbol{\beta}_0) \{e_t(\boldsymbol{\beta}_0) - \epsilon_t(\boldsymbol{\beta}_0)\}, \\
\mathbf{S}_{13n} &= n^{-1} \sum_{t=1}^{n} \mathbf{c}_t(\boldsymbol{\beta}_0) \epsilon_t(\boldsymbol{\beta}_0).
\end{aligned}$$

For the term $\mathbf{S}_{13n}$, an application of the ergodic Theorem gives the following limit:

$$\mathbf{S}_{13n} \xrightarrow{a.s.} E\{\mathbf{c}_t(\boldsymbol{\beta}_0)\epsilon_t(\boldsymbol{\beta}_0)\} = \mathbf{0}.$$

Now we consider the ith component of $\mathbf{S}_{11n}$:

$$S_{11in} = n^{-1} \sum_{t=1}^{n} \{d_{t,i}(\boldsymbol{\beta}_0) - c_{t,i}(\boldsymbol{\beta}_0)\} e_t(\boldsymbol{\beta}_0).$$

The Cauchy-Schwarz inequality gives:

$$|S_{11in}| \leq \left\{ n^{-1} \sum_{t=1}^{n} \left(d_{t,i}(\boldsymbol{\beta}_0) - c_{t,i}(\boldsymbol{\beta}_0)\right)^2 \right\}^{\frac{1}{2}} \left\{ n^{-1} \sum_{t=1}^{n} e_t^2(\boldsymbol{\beta}_0) \right\}^{\frac{1}{2}}.$$

An application of (20) gives $S_{11in} \xrightarrow{a.s.} 0$ since $\boldsymbol{\beta}_{0(\mu)} \in \mathcal{C}_\delta$ and $n^{-1} \sum_{t=1}^{n} e_t^2(\boldsymbol{\beta}_0) \xrightarrow{a.s.} \sigma_0^2$. Similar arguments give $\mathbf{S}_{12n} \xrightarrow{a.s.} \mathbf{0}$. Thus $\mathbf{S}_{1n} \xrightarrow{a.s.} \mathbf{0}$.

We now consider the term $\mathbf{S}_{2n}$. We can write it as $\mathbf{S}_{2n} = \mathbf{S}_{21n} + \mathbf{S}_{22n} + \mathbf{S}_{23n}$, where

$$\begin{aligned}
\mathbf{S}_{21n} &= n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\boldsymbol{\beta}_0) \left\{ e_t(\tilde{\boldsymbol{\beta}}_n) - \epsilon_t(\tilde{\boldsymbol{\beta}}_n) \right\}, \\
\mathbf{S}_{22n} &= n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\boldsymbol{\beta}_0) \left\{ \epsilon_t(\tilde{\boldsymbol{\beta}}_n) - \epsilon_t(\boldsymbol{\beta}_0) \right\}, \\
\mathbf{S}_{23n} &= n^{-1} \sum_{t=1}^{n} \mathbf{d}_t(\boldsymbol{\beta}_0) \{\epsilon_t(\boldsymbol{\beta}_0) - e_t(\boldsymbol{\beta}_0)\}.
\end{aligned}$$

Consider $\mathbf{S}_{21n}$. Using the difference (18), the jth component of $\mathbf{S}_{21n}$ can be written as:

$$S_{21jn} = -n^{-1} \sum_{t=1}^{n} \sum_{i \geq 0} d_{t,j}(\boldsymbol{\beta}_0) \pi_{i+t}(\tilde{\boldsymbol{\beta}}_{n(\mu)})(Y_{-i} - \tilde{\mu}_n).$$

Thus the Minkovski inequality offers the following bound:

$$\begin{aligned}|S_{21jn}| &\leq \frac{K}{n}\sum_{t=1}^{n}\sum_{i\geq 0}|d_{t,j}(\boldsymbol{\beta}_0)|\rho^{i+t}\{|Y_{-i}|+|\tilde{\mu}_n|\},\\ &= \frac{K}{n}\left[\left\{\sum_{t=1}^{n}\sum_{i\geq 0}\rho^{i+t}|d_{t,j}(\boldsymbol{\beta}_0)||Y_{-i}|\right\}+|\tilde{\mu}_n|\sum_{t=1}^{n}\sum_{i\geq 0}\rho^{i+t}|d_{t,j}(\boldsymbol{\beta}_0)|\right].\end{aligned}$$

Using similar arguments that before and Assumption B(i) for the estimation of μ, that is $\tilde{\mu}_n \xrightarrow{a.s.} \mu_0$, this shows that $\mathbf{S}_{21jn} \xrightarrow{a.s.} 0$, since $\sum_{i\geq 0}\rho^i|Y_{-i}|$ is a well-defined random variable, and

$$n^{-1}\sum_{t=1}^{n}\rho^t|d_{t,j}(\boldsymbol{\beta}_0)| \leq n^{-1}\sum_{t=1}^{n}\rho^t|d_{t,j}(\boldsymbol{\beta}_0)-c_{t,j}(\boldsymbol{\beta}_0)| + n^{-1}\sum_{t=1}^{n}\rho^t|c_{t,j}(\boldsymbol{\beta}_0)| \xrightarrow{a.s.} 0.$$

This implies $\mathbf{S}_{21n} \xrightarrow{a.s.} \mathbf{0}$. Similar arguments demonstrate that $\mathbf{S}_{23n} \xrightarrow{a.s.} \mathbf{0}$. Finally, we show that $\mathbf{S}_{22n} \xrightarrow{a.s.} \mathbf{0}$. Using Lemma 1 and formula (15), we decompose $\mathbf{S}_{22n}$ as $\mathbf{S}_{22n} = \mathbf{S}_{221n} + \mathbf{S}_{222n} + \mathbf{S}_{223n} + \mathbf{S}_{224n}$, where:

$$\begin{aligned}\mathbf{S}_{221n} &= (\mu_0-\tilde{\mu}_n)\left\{n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0)\right\},\\ \mathbf{S}_{222n} &= n^{-1}\sum_{t=1}^{n}\left[\mathbf{d}_t(\boldsymbol{\beta}_0)\sum_{i\geq 1}\left\{\pi_i(\tilde{\boldsymbol{\beta}}_{n(\mu)})-\pi_i(\boldsymbol{\beta}_{0(\mu)})\right\}Y_{t-i}\right],\\ \mathbf{S}_{223n} &= \left\{(\mu_0-\tilde{\mu}_n)\sum_{i\geq 1}\pi_i(\boldsymbol{\beta}_{0(\mu)})\right\}\left\{n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0)\right\},\\ \mathbf{S}_{224n} &= \left[\tilde{\mu}_n\sum_{i\geq 1}\left\{\pi_i(\boldsymbol{\beta}_{0(\mu)})-\pi_i(\tilde{\boldsymbol{\beta}}_{n(\mu)})\right\}\right]\left\{n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0)\right\}.\end{aligned}$$

Noting that $n^{-1}\sum_{t=1}^{n}d_{t,i}(\boldsymbol{\beta}_0) = n^{-1}\sum_{t=1}^{n}\{d_{t,i}(\boldsymbol{\beta}_0)-c_{t,i}(\boldsymbol{\beta}_0)\} + n^{-1}\sum_{t=1}^{n}c_{t,i}(\boldsymbol{\beta}_0)$, using expression (20) and results on Cezàro means give $n^{-1}\sum_{t=1}^{n}\{d_{t,i}(\boldsymbol{\beta}_0)-c_{t,i}(\boldsymbol{\beta}_0)\} \xrightarrow{a.s.} 0$. An application of the ergodic Theorem yields $n^{-1}\sum_{t=1}^{n}c_{t,i}(\boldsymbol{\beta}_0) \xrightarrow{a.s.} E\{c_{t,i}(\boldsymbol{\beta}_0)\}$. Thus $n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0) \xrightarrow{a.s.} E\{\mathbf{c}_t(\boldsymbol{\beta}_0)\}$. This implies that $\mathbf{S}_{221n} \xrightarrow{a.s.} \mathbf{0}$ and $\mathbf{S}_{223n} \xrightarrow{a.s.} \mathbf{0}$, invoking Assumption B(i) implying that $\tilde{\mu}_n \xrightarrow{a.s.} \mu_0$. Now we consider $\mathbf{S}_{224n}$. A first-order Taylor expansion gives:

$$\pi_i(\tilde{\boldsymbol{\beta}}_{n(\mu)}) = \pi_i(\boldsymbol{\beta}_{0(\mu)}) + \left\{\frac{\partial\pi_i(\overline{\boldsymbol{\beta}}_{(\mu)})}{\partial\boldsymbol{\beta}_{(\mu)}}\right\}^{\top}\left\{\tilde{\boldsymbol{\beta}}_{n(\mu)}-\boldsymbol{\beta}_{0(\mu)}\right\}.$$

Straightforward algebra gives the relations:

$$\begin{aligned}\mathbf{S}_{224n} &= \left[\tilde{\mu}_n \sum_{i\geq 0}\left\{\pi_i(\boldsymbol{\beta}_{0(\mu)}) - \pi_i(\tilde{\boldsymbol{\beta}}_{n(\mu)})\right\}\right] n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0), \\ &= \left[\tilde{\mu}_n \sum_{i\geq 0}\left\{\frac{\partial \pi_i(\overline{\boldsymbol{\beta}}_{(\mu)})}{\partial \boldsymbol{\beta}_{(\mu)}}\right\}^{\top}\left\{\tilde{\boldsymbol{\beta}}_{n(\mu)} - \boldsymbol{\beta}_{0(\mu)}\right\}\right] n^{-1}\sum_{t=1}^{n}\mathbf{d}_t(\boldsymbol{\beta}_0).\end{aligned}$$

Using the inequality:

$$\sup_{\boldsymbol{\beta}_{(\mu)}\in\mathcal{C}_\delta}\left|\frac{\partial \pi_i(\boldsymbol{\beta}_{(\mu)})}{\partial \boldsymbol{\beta}_{(\mu)}}\right| \leq K\rho^i,$$

for $K > 0$ and $\rho \in (0,1)$, similar arguments allow us to show $\mathbf{S}_{222n} \xrightarrow{a.s.} \mathbf{0}$. Similarly, $\mathbf{S}_{224n} \xrightarrow{a.s.} \mathbf{0}$. Collecting the results, this shows that $\mathbf{S}_{22n} \xrightarrow{a.s.} \mathbf{0}$. Similar arguments allows us to conclude that $\mathbf{S}_{3n} \xrightarrow{a.s.} \mathbf{0}$. Finally, we obtain the first part of the announced result, that is $n^{-1}\mathbf{D}_n^{\top}\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n) \xrightarrow{a.s.} \mathbf{0}$.

Brockwell and Davis (1991) showed in detail that $n^{-1}\mathbf{D}_n^{\top}\mathbf{D}_n \xrightarrow{P} \sigma_0^2\mathbf{V}^{-1}(\boldsymbol{\beta}_0)$. To show the stronger result that $n^{-1}\mathbf{D}_n^{\top}\mathbf{D}_n \xrightarrow{a.s.} \sigma_0^2\mathbf{V}^{-1}(\boldsymbol{\beta}_0)$, we use the following decomposition:

$$\mathbf{x}\mathbf{x}^{\top} = (\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^{\top} + (\mathbf{x}-\mathbf{y})\mathbf{y}^{\top} + \mathbf{y}(\mathbf{x}-\mathbf{y})^{\top} + \mathbf{y}\mathbf{y}^{\top},$$

which is applied on:

$$\begin{aligned}n^{-1}\mathbf{D}_n^{\top}\mathbf{D}_n &= n^{-1}\sum_{t=1}^{n}\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n) + \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n) + \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}^{\top}, \\ &= \mathbf{S}_{4n} + \mathbf{S}_{5n} + \mathbf{S}_{6n} + \mathbf{S}_{7n},\end{aligned}$$

where

$$\begin{aligned}\mathbf{S}_{4n} &= n^{-1}\sum_{t=1}^{n}\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}^{\top}, \\ \mathbf{S}_{5n} &= n^{-1}\sum_{t=1}^{n}\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}\mathbf{c}_t^{\top}(\tilde{\boldsymbol{\beta}}_n), \\ \mathbf{S}_{6n} &= n^{-1}\sum_{t=1}^{n}\mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\{\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\}^{\top}, \\ \mathbf{S}_{7n} &= n^{-1}\sum_{t=1}^{n}\mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)\mathbf{c}_t^{\top}(\tilde{\boldsymbol{\beta}}_n).\end{aligned}$$

The almost sure limit of $n^{-1}\mathbf{D}_n^{\top}\mathbf{D}_n$ comes from the study of the expression $n^{-1}\sum_{t=1}^{n}\mathbf{c}_t(\boldsymbol{\beta}_0)\mathbf{c}_t^{\top}(\boldsymbol{\beta}_0)$, which by the ergodic Theorem converges to the following limit:

$$n^{-1}\sum_{t=1}^{n}\mathbf{c}_t(\boldsymbol{\beta}_0)\mathbf{c}_t^{\top}(\boldsymbol{\beta}_0) \xrightarrow{a.s.} E\left\{\mathbf{c}_t(\boldsymbol{\beta}_0)\mathbf{c}_t^{\top}(\boldsymbol{\beta}_0)\right\} = \sigma_0^2\mathbf{V}^{-1}(\boldsymbol{\beta}_0).$$

The argument leading to the last equality is justified in Brockwell and Davis (1991, p. 268). Note that $\mathbf{S}_{4n}$, $\mathbf{S}_{5n}$ and $\mathbf{S}_{6n}$ involve the differences $\mathbf{d}_t(\tilde{\boldsymbol{\beta}}_n) - \mathbf{c}_t(\tilde{\boldsymbol{\beta}}_n)$ and by hypothesis $\tilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$. Using similar and lengthy arguments similar to those called upon in the study of $n^{-1}\mathbf{D}_n^\top \mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n)$, $\mathbf{S}_{4n} \xrightarrow{a.s.} \mathbf{0}$, $\mathbf{S}_{5n} \xrightarrow{a.s.} \mathbf{0}$ and $\mathbf{S}_{6n} \xrightarrow{a.s.} \mathbf{0}$ and the announced result is proved:

$$n^{-1}\mathbf{D}_n^\top \mathbf{D}_n \xrightarrow{a.s.} \sigma_0^2 \mathbf{V}^{-1}(\boldsymbol{\beta}_0).$$

Thus $\hat{\boldsymbol{\beta}}_{1SGN,n} \xrightarrow{a.s.} \boldsymbol{\beta}_0$. Concerning the study of $\hat{\sigma}^2_{1SGN,n}$, note that:

$$n^{-1}\sum_{t=1}^{n} e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - n^{-1}\sum_{t=1}^{n} e_t^2(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0,$$

and $n^{-1}\sum_{t=1}^{n} e_t^2(\boldsymbol{\beta}_0) - n^{-1}\sum_{t=1}^{n} \epsilon_t^2(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0$. An application of the ergodic Theorem gives the almost sure convergence $n^{-1}\sum_{t=1}^{n} \epsilon_t^2 \xrightarrow{a.s.} E(\epsilon_t^2) = \sigma_0^2$, and consequently:

$$\hat{\sigma}^2_{1SGN,n} \xrightarrow{a.s.} \sigma_0^2.$$

This shows the almost sure convergence of $\hat{\boldsymbol{\nu}}_{1SGN,n}$ to $\boldsymbol{\nu}_0$. This completes the proof of the Theorem. □

4. Asymptotic normality

Brockwell and Davis (1991) provided a sketch of the proof establishing that one-step Gauss-Newton estimators and maximum Gaussian likelihood estimators share the same asymptotic distribution in ARMA models. Here, we present a detailed proof of this claim, by showing that the first-order term of the asymptotically equivalent expression is the same that the one of the (exact) maximum likelihood estimator. First-order terms may be of interest in their own rights, when one concentrates on efficient estimation in ARMA models, which is made possible using modern software. This was the route taken in Duchesne et al. (2016). The first-order expansion is given in the next Theorem.

Theorem 2. *Let $\{Y_t, t \in \mathbb{Z}\}$ be a stochastic difference equation generated by (1). Consider the one-step Gauss-Newton estimator defined by (11) and (12). Under Assumptions A and B, the following expansion is valid:*

$$n^{1/2}(\hat{\boldsymbol{\nu}}_{1SGN,n} - \boldsymbol{\nu}_0) = n^{-1/2}\sum_{t=1}^{n} \begin{pmatrix} \sigma_0^{-2}\mathbf{V}(\boldsymbol{\beta}_0)\mathbf{W}_t\epsilon_t \\ \epsilon_t^2 - \sigma_0^2 \end{pmatrix} + \mathbf{o}_p(1),$$

where $\mathbf{W}_t = (W_{t,1}, \ldots, W_{t,p+q+1})^\top$, with, for $t \in \mathcal{I}_n$,

$$W_{t,i} = \begin{cases} \Phi_p^{-1}(\boldsymbol{\phi}_0, B)\epsilon_{t-i}(\boldsymbol{\beta}_0)\mathbb{1}_{\{t-i\in\mathcal{I}_n\}}, & \text{for } i = 1, \ldots, p, \\ \Theta_q^{-1}(\boldsymbol{\theta}_0, B)\epsilon_{t+p-i}(\boldsymbol{\beta}_0)\mathbb{1}_{\{t+p-i\in\mathcal{I}_n\}}, & \text{for } i = p+1, \ldots, p+q, \\ c_0, & \text{for } i = p+q+1, \end{cases}$$

where $c_0 = \Phi_p(\boldsymbol{\phi}_0, 1)/\Theta_q(\boldsymbol{\theta}_0, 1)$ and $\mathbf{V}(\boldsymbol{\beta}_0) = \sigma_0^2\{E(\mathbf{W}_1\mathbf{W}_1^\top)\}^{-1}$.

Proof. We first show that

$$n^{1/2}\left(\hat{\boldsymbol{\beta}}_{1SGN,n}-\boldsymbol{\beta}_0\right)=n^{-1/2}\sum_{t=1}^{n}\left\{\sigma_0^{-2}\mathbf{V}(\boldsymbol{\beta}_0)\mathbf{W}_t\epsilon_t\right\}+\mathbf{o}_P(1), \tag{21}$$

as $n\to\infty$. Using the expression (10), the one-step Gauss-Newton estimator can be written as:

$$\begin{aligned}
n^{1/2}(\hat{\boldsymbol{\beta}}_{1SGN,n}-\boldsymbol{\beta}_0) &= n^{1/2}(\tilde{\boldsymbol{\beta}}_n+\widehat{\boldsymbol{\Delta}\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0),\\
&= n^{1/2}\{\tilde{\boldsymbol{\beta}}_n+(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{e}_n(\tilde{\boldsymbol{\beta}}_n)-\boldsymbol{\beta}_0\},\\
&= n^{1/2}(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{e}_n(\boldsymbol{\beta}_0)-n^{1/2}(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\mathbf{D}_n^\top\mathbf{Z}_n,\\
&= (n^{-1}\mathbf{D}_n^\top\mathbf{D}_n)^{-1}n^{-1/2}\mathbf{D}_n^\top\mathbf{e}_n(\boldsymbol{\beta}_0)-(n^{-1}\mathbf{D}_n^\top\mathbf{D}_n)^{-1}n^{-1/2}\mathbf{D}_n^\top\mathbf{Z}_n.
\end{aligned}$$

According to Theorem 1, $n^{-1}\mathbf{D}_n^\top\mathbf{D}_n\xrightarrow{a.s.}\sigma_0^2\mathbf{V}^{-1}(\boldsymbol{\beta}_0)$, which implies that $n^{-1}\mathbf{D}_n^\top\mathbf{D}_n\xrightarrow{P}\sigma_0^2\mathbf{V}^{-1}(\boldsymbol{\beta}_0)$. In order to prove (21), it is sufficient to prove:

$$n^{-1/2}\mathbf{D}_n^\top\mathbf{e}_n(\boldsymbol{\beta}_0) \equiv n^{-1/2}\mathbf{T}_n=n^{-1/2}\sum_{t=1}^{n}\mathbf{W}_t\epsilon_t+\mathbf{o}_P(1), \tag{22}$$

$$n^{-1/2}\mathbf{D}_n^\top\mathbf{Z}_n \equiv n^{-1/2}\mathbf{U}_n=\mathbf{o}_P(1). \tag{23}$$

First, we show expression (22). The components of the $(p+q+1)\times 1$ random vector $\mathbf{T}_n$ satisfies $\mathbf{T}_n=(T_{1,n},\ldots,T_{p+q+1,n})^\top$. The ith component of $\mathbf{T}_n$ is given by:

$$n^{-1/2}T_{i,n}=n^{-1/2}\sum_{t=1}^{n}d_{t,i}(\tilde{\boldsymbol{\beta}}_n)e_t(\boldsymbol{\beta}_0).$$

Only $d_{t,i}(\tilde{\boldsymbol{\beta}}_n)$ is function of the preliminary estimator $\tilde{\boldsymbol{\beta}}_n$. A first-order Taylor expansion of $d_{t,i}(\tilde{\boldsymbol{\beta}}_n)$ around the point $\boldsymbol{\beta}_0$ yields:

$$d_{t,i}(\tilde{\boldsymbol{\beta}}_n)=d_{t,i}(\boldsymbol{\beta}_0)+(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)^\top\frac{\partial d_{t,i}(\bar{\boldsymbol{\beta}}_{tn})}{\partial\boldsymbol{\beta}},$$

where $\bar{\boldsymbol{\beta}}_{tn}$ is between $\tilde{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. However, one easily shows that:

$$\left\|\frac{\partial}{\partial\beta_j}\left\{d_{t,i}(\boldsymbol{\beta}_{tn}^*)\right\}\right\|\leq K.$$

From Assumption B(ii), the preliminary estimator of $\boldsymbol{\beta}_0$ satisfies the rate $\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0=\mathbf{o}_P(n^{-1/4})$. Consequently, the following expansion is obtained:

$$n^{-1/2}\sum_{t=1}^{n}d_{t,i}(\tilde{\boldsymbol{\beta}}_n)\epsilon_t(\boldsymbol{\beta}_0)-n^{-1/2}\sum_{t=1}^{n}d_{t,i}(\boldsymbol{\beta}_0)\epsilon_t(\boldsymbol{\beta}_0)=o_P(1).$$

Proceeding as in the proof of the almost sure convergence of $\hat{\boldsymbol{\nu}}_{1SGN,n}$, it is relatively easy to show that the residuals can be replaced by the innovation process $\{\epsilon_t(\cdot)\}$ at $\boldsymbol{\beta}_0$:

$$n^{-1/2}\sum_{t=1}^{n} d_{t,i}(\boldsymbol{\beta}_0)\epsilon_t(\boldsymbol{\beta}_0) = n^{-1/2}\sum_{t=1}^{n} W_{t,i}\epsilon_t + \mathbf{o}_P(1).$$

This shows the result for the estimators of the autoregressive and moving average parts. Now, it remains to study the variance estimator. Consider the following decomposition:

$$\begin{aligned} n^{1/2}(\hat{\sigma}^2_{1SGN,n} - \sigma_0^2) &= n^{-1/2}\sum_{t=1}^{n}\{e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \sigma_0^2\}, \\ &= n^{-1/2}\sum_{t=1}^{n}\{e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t^2(\boldsymbol{\beta}_0)\} + n^{-1/2}\sum_{t=1}^{n}\left\{\epsilon_t^2(\boldsymbol{\beta}_0) - \sigma_0^2\right\}. \end{aligned}$$

For $t \in I_n$, a simple application of the Cauchy-Schwarz inequality and the Minkowski inequality give the following bound:

$$\begin{aligned} \|e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t^2\| &= \|\{e_t(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t\}\{e_t(\hat{\boldsymbol{\beta}}_{1SGN,n}) + \epsilon_t\}\|, \\ &\leq \|e_t(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t\|\ \|e_t(\hat{\boldsymbol{\beta}}_{1SGN,n}) + \epsilon_t\|, \\ &\leq K\|Y_1 - \mu\|(1-\rho)^{-1}\rho^t(\|e_t(\hat{\boldsymbol{\beta}}_{1SGN,n})\| + \|\epsilon_t\|), \\ &\leq K\rho^t. \end{aligned}$$

Thus

$$\|n^{-1/2}\sum_{t=1}^{n}\{e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t^2\}\| \leq n^{-1/2}K(1-\rho)^{-1}(1-\rho^n) = O(n^{-1/2}).$$

Consequently, using Brockwell and Davis (1991, Proposition 6.2.3), this shows:

$$n^{-1/2}\sum_{t=1}^{n}\{e_t^2(\hat{\boldsymbol{\beta}}_{1SGN,n}) - \epsilon_t^2\} = O_P(n^{-1/2}) = o_P(1).$$

This shows the announced result for the variance estimator. Collecting the results, this shows (22). We now study the expression given by (23). As for (22), it is enough to study $n^{-1/2}\mathbf{U}_n$ component by component. Here, the components of the $(p+q+1)\times 1$ random vector $\mathbf{U}_n$ are collected in $\mathbf{U}_n = (U_{1,n}, \ldots, U_{p+q+1,n})^\top$. The kth component of $\mathbf{U}_n$, $k = 1, \ldots, p+q+1$, is given by:

$$n^{-1/2}U_{k,n} = n^{-1/2}\sum_{t=1}^{n} d_{t,k}(\tilde{\boldsymbol{\beta}}_n)\frac{1}{2}\sum_{i=1}^{p+q+1}\sum_{j=1}^{p+q+1}\frac{\partial^2 e_t}{\partial\beta_i\partial\beta_j}(\boldsymbol{\beta}^*_{tn})(\beta_{0i} - \tilde{\beta}_i)(\beta_{0j} - \tilde{\beta}_j).$$

Recall that from Assumption B(ii), the preliminary estimator satisfies $(\beta_{0i} - \tilde{\beta}_i)(\beta_{0j} - \tilde{\beta}_j) = o_P(n^{-1/2})$. Thus, it suffices to show that:

$$n^{-1}\sum_{t=1}^{n}\left|d_{t,k}(\tilde{\boldsymbol{\beta}}_n)\frac{\partial^2 e_t(\boldsymbol{\beta}^*_{tn})}{\partial\beta_i\partial\beta_j}\right| = O_P(1),$$

$i, j, k = 1, \ldots, p+q+1$. We prove that statement by showing that the norm $\|\cdot\|$ of the left member is bounded. The Cauchy-Schwarz inequality yields the following bound:

$$\left\| d_{t,k}(\tilde{\boldsymbol{\beta}}_n) \frac{\partial^2 e_t(\boldsymbol{\beta}^*_{tn})}{\partial\beta_i \partial\beta_j} \right\| \leq \left\| d_{t,k}(\tilde{\boldsymbol{\beta}}_n) \right\| \left\| \frac{\partial^2 e_t(\boldsymbol{\beta}^*_{tn})}{\partial\beta_i \partial\beta_j} \right\| \leq K.$$

By Assumption, $\tilde{\boldsymbol{\beta}}_n$ converges almost surely to $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}^*_{tn}$ is between $\boldsymbol{\beta}_0$ and $\tilde{\boldsymbol{\beta}}_n$, and so it also belongs to an open neighborhood of $\boldsymbol{\beta}_0$. This concludes the proof. □

It follows that $\hat{\boldsymbol{\beta}}_{1SGN,n}$ is asymptotically efficient when the normality assumption holds true, with an asymptotic normal distribution. Even if the error process is not Gaussian, $\hat{\boldsymbol{\beta}}_{1SGN,n}$ and the Gaussian maximum likelihood estimators share the same asymptotic normal distribution. See also the discussions in Brockwell and Davis (1991, p. 265) and Brockwell and Davis (2002, p. 159). Adopting the $\mathcal{AN}$ notation as in Serfling (1980), this means that:

$$\hat{\boldsymbol{\beta}}_{1SGN,n} \text{ is } \mathcal{AN}_{p+q+1}(\boldsymbol{\beta}_0, n^{-1}\mathbf{V}(\boldsymbol{\beta}_0)),$$

with $\mathbf{V}(\boldsymbol{\beta}_0)$ as described in Theorem 2. Note that $\mathbf{V}(\boldsymbol{\beta}_0)$ can be written as a block diagonal matrix. We write $\mathbf{V}(\boldsymbol{\beta}_0) = \sigma_0^2 \text{diag}(\mathbf{A}^{-1}(\boldsymbol{\beta}_0), c_0^{-2})$, where the $(p+q)\times(p+q)$ matrix $\mathbf{A}(\boldsymbol{\beta}_0)$ is given in Brockwell and Davis (1991, equ. (8.8.3), p. 258).

5. Empirical evidence

In the previous sections we studied the theoretical properties of one-step Gauss-Newton estimators. >From an empirical point-of-view, it seems relevant to inquire about the finite sample properties of that estimation method. It is also of interest to investigate the effects of the preliminary estimator on the empirical biases and variances. More particularly, it is relevant to appreciate if efficiency of the initial method is improved by using the proposed one-step methodology. Consequently, it may be useful to study the exact biases and mean squared errors of the estimators. To partially answer these questions, we conducted some Monte Carlo experiments.

The following MA(1) model has been used:

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}, \tag{24}$$

where $\{\epsilon_t\}$ is a strong white noise process. To consider an MA(1) model is motivated by the fact that the moment estimator of θ yields an inefficient procedure, but it is still often recommended as a preliminary estimation method in authoritative textbooks, see, e.g., Brockwell and Davis (1991, p. 254).

Since estimating the mean parameter μ by the sample mean gives an asymptotically efficient method, see, e.g., Brockwell and Davis (1991, p. 220), we considered using one-step Gauss-Newton estimators of θ using mean-corrected data. As preliminary estimators, we

used the moment estimator of θ as described in Brockwell and Davis (1991, p. 253), and also the estimators based on the innovations algorithm (see Brockwell and Davis (1991, p. 245)). We denote these estimators $\hat{\theta}_{MM|\bar{X},n}$ and $\hat{\theta}_{IA|\bar{X},n}$, respectively. When $\{\epsilon_t\}$ is strong white noise, such that $E(\epsilon_t^4) < \infty$, the asymptotic distributions of these estimators are

$$\hat{\theta}_{MM|\bar{X},n} \text{ is } \mathcal{AN}\left(\theta, n^{-1}\frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{(1-\theta^2)^2}\right),$$

and

$$\hat{\theta}_{IA|\bar{X},n} \text{ is } \mathcal{AN}\left(\theta, n^{-1}\right),$$

respectively. See Brockwell and Davis (1991, pp. 253-254). In addition, the MLE of $(\theta, \mu)^\top$ is denoted $(\hat{\theta}_{MLE,n}, \hat{\mu}_{MLE,n})^\top$. It is well-known that

$$\hat{\theta}_{MLE,n} \text{ is } \mathcal{AN}\left(\theta, n^{-1}(1-\theta^2)\right).$$

We considered the one-step Gauss-Newton estimator of θ based on mean-corrected observations, using $\hat{\theta}_{MM|\bar{X},n}$ and $\hat{\theta}_{IA|\bar{X},n}$ as preliminary estimators. We denote these estimators $\hat{\theta}_{1SGN|\bar{X},n}(\hat{\theta}_{MM|\bar{X},n})$ and $\hat{\theta}_{1SGN|\bar{X},n}(\hat{\theta}_{IA|\bar{X},n})$, respectively. We also considered one-step Gauss-Newton estimators of the vector $\boldsymbol{\beta} = (\theta, \mu)^\top$, based on $(\hat{\theta}_{MM|\bar{X},n}, \bar{X}_n)^\top$ and $(\hat{\theta}_{IA|\bar{X},n}, \bar{X}_n)^\top$, denoted $\hat{\boldsymbol{\beta}}_{1SGN}(\hat{\theta}_{MM|\bar{X},n}, \bar{X}_n)$ and $\hat{\boldsymbol{\beta}}_{1SGN}(\hat{\theta}_{IA|\bar{X},n}, \bar{X}_n)$, respectively. More precisely, the estimators of θ are noted $\hat{\theta}_{1SGN}(\hat{\theta}_{MM|\bar{X},n}, \bar{X}_n)$ and $\hat{\theta}_{1SGN}(\hat{\theta}_{IA|\bar{X},n}, \bar{X}_n)$, respectively. The asymptotic distribution of all these one-step estimators of θ is the same that the one of the MLE: $\mathcal{AN}\left(\theta, n^{-1}(1-\theta^2)\right)$.

We considered the MA(1) process with $\theta \in \{\pm 0.1, \pm 0.5, \pm 0.7, \pm 0.9\}$. The mean of the process was set to $E(Y_t) = 1$. The white noise process $\{\epsilon_t\}$ was assumed Gaussian, such that $\epsilon_t \sim \mathcal{N}(0,1)$. We considered two sample sizes: $n = 50, 100$.

Based on an estimator $\hat{\theta}_n$, the Monte Carlo mean, relative bias, variance and mean squared errors have been computed according to the formula:

$$\begin{aligned}
E_{MC}(\hat{\theta}) &= N_{sim}^{-1}\sum_{i=1}^{N_{sim}}\hat{\theta}^{(i)}, \\
RB_{MC}(\hat{\theta}) &= \left\{E_{MC}(\hat{\theta}) - \theta\right\}/\theta, \\
V_{MC}(\hat{\theta}) &= N_{sim}^{-1}\sum_{i=1}^{N_{sim}}\left\{\hat{\theta}^{(i)} - E_{MC}(\hat{\theta})\right\}^2, \\
MSE_{MC}(\hat{\theta}) &= N_{sim}^{-1}\sum_{i=1}^{N_{sim}}\left\{\hat{\theta}^{(i)} - \theta\right\}^2,
\end{aligned}$$

where $\hat{\theta}^{(i)}$ corresponds to the value of the estimator at iteration i, $i = 1, \ldots, N_{sim}$. In our simulation experiments, we considered $N_{sim} = 5000$. All the computer code has been written with the **R** language.

Table 1: Empirical relative biases and mean squared errors of μ in the MA(1) model when $\theta \in \{\pm 0.9, \pm 0.7, \pm 0.5, \pm 0.1\}$. The number of replications is $N_{sim} = 5000$.

	$\theta=-0.9$		$\theta=-0.7$		$\theta=-0.5$		$\theta=-0.1$		$\theta=0.1$		$\theta=0.5$		$\theta=0.7$		$\theta=0.9$	
	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE
$n=50$																
$\bar{X}_n$	-0.001	0.001	0.000	0.002	-0.002	0.005	0	0.016	-0.001	0.025	0.001	0.045	-0.002	0.057	0.007	0.072
$\hat{\mu}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n}, \bar{X})$	0.000	0.006	-0.002	0.006	-0.002	0.007	0	0.016	-0.001	0.025	0.003	0.092	-0.001	0.137	0.002	0.131
$\hat{\mu}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n}, \bar{X})$	0.000	0.002	-0.001	0.002	-0.002	0.005	0	0.016	-0.001	0.025	0.001	0.045	-0.004	0.066	0.008	0.111
$\hat{\mu}_{MLE}$	0.000	0.000	0.000	0.002	-0.002	0.005	0	0.016	-0.001	0.025	0.001	0.045	-0.002	0.057	0.007	0.071
$n=100$																
$\bar{X}_n$	0	0.000	0.000	0.001	0.001	0.003	0.001	0.008	-0.001	0.012	0.002	0.024	0.001	0.029	-0.003	0.036
$\hat{\mu}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n}, \bar{X})$	0	0.002	-0.001	0.002	0.001	0.003	0.001	0.008	-0.001	0.012	0.001	0.082	-0.002	0.142	-0.003	0.111
$\hat{\mu}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n}, \bar{X})$	0	0.000	0.000	0.001	0.001	0.003	0.001	0.008	-0.001	0.012	0.002	0.024	0.000	0.029	-0.005	0.078
$\hat{\mu}_{MLE}$	0	0.000	0.000	0.001	0.001	0.003	0.001	0.008	-0.001	0.012	0.002	0.024	0.001	0.029	-0.002	0.036

Table 2: Empirical relative biases and mean squared errors of θ in the MA(1) model for $\theta \in \{\pm 0.9, \pm 0.7, \pm 0.5, \pm 0.1\}$. The number of replications is $N_{sim} = 5000$.

	$\theta=-0.9$		$\theta=-0.7$		$\theta=-0.5$		$\theta=-0.1$		$\theta=0.1$		$\theta=0.5$		$\theta=0.7$		$\theta=0.9$	
	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE	RB	MSE
$n=50$																
$\hat{\theta}_{MM\|\bar{X},n}$	0.152	0.085	0.013	0.066	-0.037	0.063	-0.023	0.024	-0.022	0.022	-0.016	0.058	-0.071	0.072	-0.200	0.106
$\hat{\theta}_{IA\|\bar{X},n}$	0.109	0.037	0.033	0.025	0.000	0.025	-0.028	0.025	-0.033	0.025	-0.054	0.028	-0.081	0.031	-0.148	0.047
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n})$	0.009	0.035	-0.087	0.053	-0.080	0.060	-0.025	0.025	-0.024	0.023	0.016	0.043	0.019	0.036	-0.060	0.032
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n})$	0.011	0.022	-0.028	0.023	-0.033	0.025	-0.029	0.026	-0.024	0.025	-0.020	0.021	-0.021	0.018	-0.054	0.020
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n}, \bar{X})$	0.036	0.025	-0.068	0.040	-0.073	0.053	-0.024	0.025	-0.024	0.023	0.015	0.043	0.018	0.036	-0.063	0.032
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n}, \bar{X})$	0.039	0.017	-0.018	0.019	-0.031	0.024	-0.029	0.026	-0.024	0.025	-0.020	0.021	-0.022	0.018	-0.056	0.020
$\hat{\theta}_{MLE}$	-0.071	0.011	-0.091	0.034	-0.067	0.037	-0.035	0.030	-0.019	0.028	0.001	0.022	0.013	0.017	0.011	0.008
$n=100$																
$\hat{\theta}_{MM\|\bar{X},n}$	0.110	0.059	-0.018	0.050	-0.035	0.041	-0.011	0.011	-0.008	0.010	0.001	0.035	-0.017	0.050	-0.135	0.067
$\hat{\theta}_{IA\|\bar{X},n}$	0.064	0.017	0.016	0.012	0.000	0.012	-0.013	0.012	-0.012	0.011	-0.027	0.012	-0.040	0.013	-0.082	0.019
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n})$	-0.012	0.016	-0.089	0.036	-0.053	0.032	-0.011	0.011	-0.008	0.010	0.017	0.024	0.051	0.027	-0.012	0.014
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n})$	0.001	0.008	-0.014	0.008	-0.014	0.010	-0.013	0.011	-0.007	0.011	-0.010	0.009	-0.010	0.007	-0.021	0.007
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{MM\|\bar{X},n}, \bar{X})$	-0.001	0.013	-0.083	0.032	-0.051	0.031	-0.011	0.011	-0.008	0.010	0.017	0.024	0.051	0.027	-0.013	0.014
$\hat{\theta}_{1SGN\|\bar{X},n}(\hat{\theta}_{IA\|\bar{X},n}, \bar{X})$	0.012	0.007	-0.013	0.008	-0.013	0.010	-0.013	0.011	-0.007	0.011	-0.010	0.009	-0.010	0.007	-0.022	0.007
$\hat{\theta}_{MLE}$	-0.053	0.007	-0.035	0.010	-0.025	0.011	-0.015	0.012	-0.005	0.011	0.000	0.009	0.005	0.007	0.012	0.003

The results are given in Tables 1 and 2. Concerning the estimation of μ, except for large values of θ, all the biases were small and the MSE were similar. Joint estimation of μ and θ by the one-step Gauss-Newton estimators yielded a somewhat inferior strategy. Since the sample mean as a preliminary estimator of μ was asymptotically efficient, it seems that for large θ an unnecessary noise was created using the one-step method. When $n = 50$, this suggests that it may be better to use mean-corrected observations. For $n = 100$, efficiency improved, and when θ was estimated using the innovations algorithm, only the case $\theta = 0.9$ offered a substantial difference compared to the MLE.

We now discuss the estimation of θ. In several cases, empirical biases using the method of moments or the innovations algorithm were substantial. As expected, the MSE of these methods were larger than the one of the MLE, since these estimators are known

inefficient. To consider 1SGN estimators improved significantly the efficiency. Performing a one-step iteration, to estimate θ based on mean-corrected observations or performing joint estimation yielded very similar results. However, the impact of the preliminary estimator remains non-negligible. >From our simulation results, to use a preliminary estimator of θ based on the innovations algorithm offered better efficiency results. Except for $\theta = 0.9$, the MSE of 1SGN estimators were similar to those of the MLE, particularly when $n = 100$.

6. Discussion and conclusion

In this note, we provided a proof of the strong consistency and we also studied the asymptotic normality of one-step Gauss-Newton estimators for ARMA time series models. In a small simulation study, we illustrated the empirical properties of the estimators when the preliminary estimator is obtained from the method of moments, and from the innovations algorithm. In a simple moving-average model, we found that it may be reasonable to use the sample mean as a preliminary estimator of μ and performing a one-step estimator of the moving-average parameter. To consider one-step Gauss-Newton estimators improved the efficiency, and the choice of the preliminary estimator may have an impact in finite samples. In our simulations, to consider a preliminary estimator based on the innovations algorithm was better than to use an estimator based on the method of moments.

7. Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

References

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. 2nd Edition. Springer-Verlag, New York.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. 2nd Edition. Springer Texts in Statistics. Springer-Verlag, New York.

Duchesne, P., Lafaye de Micheaux, P. and Tagne Tatsinkou, J. F. (2016). 'Estimating the mean and its effects on Neyman smooth tests of normality for ARMA models'. *The Canadian Journal of Statistics* **44**, 241–270.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

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

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

Fuller, W. A. (1996). *Introduction to Statistical Time Series*. 2nd Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

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

Luceño, A. (1993). 'A fast algorithm for the repeated evaluation of the likelihood of a general linear process for long series'. *Journal of the American Statistical Association* **88**, 229–236.

McLeod, A. I. (1993). 'A note on ARMA model parameter redundancy'. *Journal of Time Series Analysis* **14**, 207–208.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc., New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Yao, Q. and Brockwell, P. J. (2006). 'Gaussian maximum likelihood estimation for ARMA models. I. Time series'. *Journal of Time Series Analysis* **27**, 857–875.