A New Minimum Distance Estimation Procedure of ARFIMA Processes

Laura Mayoral*
Dept. of Economics and Business
Universidad Pompeu Fabra

Fist version: February, 2001. This version: July, 2005

Abstract

A new parametric minimum distance time-domain estimator for ARFIMA processes is introduced in this paper. The proposed estimator minimizes the sum of squared correlations of residuals obtained after filtering a series through ARFIMA parameters. The estimator is easy to compute and is consistent, asymptotically normally distributed and efficient for fractionally integrated (FI) processes with an integration order d strictly greater than -0.75. Therefore, it can be applied to both stationary and non-stationary processes. Deterministic components are also allowed in the DGP. Furthermore, as a by-product, the estimation procedure provides an immediate check on the adequacy of the specified model. This is so because the criterion function, when evaluated at the estimated values, coincides with the Box-Pierce goodness of fit statistic. Empirical applications and Monte-Carlo simulations supporting the analytical results and showing the good performance of the estimator in finite samples are also provided.

Keywords: fractional integration; minimum distance estimation; long memory and non-stationary time series.

JEL Classification: C13, C22.

*I am very grateful to Juan J. Dolado and Jesús Gonzalo for their valuable comments and also to Gabor Lugosi, Francesc Mármol, Albert Satorra, Carlos Velasco, three anonimous referees, the editor and participants at ESEM 2001 (Lausanne). I acknowledge financial support from grants SEC2003-04429 and SEC2003-04476 and the Barcelona Economics Program of CREA.

1. Introduction

A new estimation procedure for Autoregressive Fractionally Integrated Moving Average (ARFIMA) processes is proposed in this paper. First introduced by Granger and Joyeux (1980) and Hosking (1981), these processes have become very popular due to their ability in providing a good characterization of the long-run properties of many economic and financial time series. They are also very useful for modeling multivariate time series, since they are able to capture a larger number of long term equilibrium relations among economic variables than the traditional multivariate ARIMA models. See Baillie (1996) and Henry and Zaffaroni (2002) for surveys on this topic.

The estimator introduced in this paper belongs to the Minimum Distance (MD) class. The idea of the estimation procedure is quite simple: the parameters of the ARFIMA model are estimated by minimizing the sum of the squared autocorrelations of the residuals, obtained after filtering the original series through ARFIMA parameters. The proposed estimator is closely related to the MD estimator considered in Tieslau, Schmidt and Baillie (1996), and to the Adjusted MD estimator proposed in Chung and Schmidt (1995). Nevertheless, as it will be seen shortly, it presents important advantages over those estimators. It is denoted "Generalized Minimum Distance" (GMD) estimator since it extends previous approaches in this area to more general setups. In particular, the proposed estimator is easy to compute, has good asymptotic and finite sample properties and is able to circumvent most of the problems present in the above-mentioned techniques. It can be applied to $FI(d)$ series for values of $d > -0.75$, thus covering stationary as well as non-stationary ranges of d. This technique has been developed in the time domain, usually preferred in applied work. Relative to other time domain approaches, such as Maximum Likelihood estimation (MLE), it presents the advantage that it is not necessary to specify a particular distribution for the innovation process. Also, it is computationally faster than exact MLE for more complex ARFIMA(p, d, q) processes, since the evaluation of the likelihood requires finding the inverse of a $T \times T$ matrix of population autocovariances, whose elements are complicated (hypergeometric) functions of the unknown parameters. For large sample sizes, accurate computation

of such matrices could be a non-trivial task, even with today's computing technology. As it will be seen, the computation of this estimator does not entail these difficulties.

Frequency domain estimators are also very popular in this literature, mainly due to their computational simplicity and their good asymptotic properties, as it is the case of the Whittle estimator (see Fox and Taqqu, (1986) and Dahlhaus, (1989)). This estimator is efficient only if the memory parameter d is known to lie in the stationary and invertible range and this restriction is correctly imposed. For the general case where d is completely unknown and possibly non-stationary, Velasco and Robinson (2000) showed that it is needed to resort to tapered data to achieve consistency. Tapering increases the variance of the estimators and therefore induces an efficiency loss. This implies that, in the general case where no information about d is available before estimation, the Whittle estimator is not efficient.

Another interesting feature is that the estimation procedure introduced in this paper provides, as a by-product, an immediate check on adequacy of the specified parametric model. This is so because the criterion function, evaluated at the estimated values, yields the Box-Pierce goodness-of-fit statistic which has been largely used for these purposes in the literature.

The rest of the paper is structured as follows. The ARFIMA model and the definition of the residuals are introduced in Section 2. The GMD estimation procedure and the asymptotic properties of the estimator are discussed in Section 3. The results of some simulation experiments, designed to evaluate the performance in finite samples of the proposed estimator, are described in Section 4. Section 5 derives the asymptotic distribution of residual autocorrelations and applies the result to the Box-Pierce (1970) and Box-Ljung (1978) goodness-of-fit statistics. Section 6 investigates their finite sample performance. An application of the described methods to empirical data is provided in Section 7. The conclusions of the paper are presented in Section 8. All proofs are gathered in Appendix 1. Appendix 2 incorporates the results of other Monte Carlo experiments.

The following conventional notation is adopted throughout the paper. L is the lag operator; $\Delta = (1-L)$; $\|.\|$ denotes the Euclidean norm for matrices and $\Gamma(.)$ the gamma function; "$\xrightarrow{w}$" and "$\xrightarrow{p}$" denote weak convergence and convergence in probability, respec-

tively; $\{\pi_i(d)\}_{i=0}^{\infty}$ represents the sequence of coefficients associated with the expansion of Δ^d in powers of L, such that $\Delta^d = \pi_0(d) + \pi_1(d)L + \pi_2(d)L^2 + ...$, and

$$\pi_i(d) = \frac{\Gamma(i-d)}{\Gamma(-d)\Gamma(i+1)}. \tag{1}$$

2. The model

The process y_t, observed at time $t = 1, ...T$, is an $ARFIMA(p, d_0, q)$ process whose memory parameter, d_0, belongs to the closed interval $[\nabla_1, \nabla_2]$, with $-0.75 < \nabla_1 < \nabla_2 < \infty$. For stationary values of d_0 ($d_0 < 1/2$), y_t can be written as,

$$\Phi_0(L)\Delta^{d_0}(y_t - \mu_0) = \Theta_0(L)\varepsilon_t,\ t = 0, \pm 1, ..., \tag{2}$$

where $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is a sequence of i.i.d. zero-mean random variables with unknown variance σ^2 and finite fourth moment, $E(\varepsilon_t^4) = \mu_4 < \infty$. $\Phi_0(L)$ and $\Theta_0(L)$ are autoregressive and moving average polynomials of order p and q, respectively, with all their roots outside the unit circle. Throughout, it will be assumed that p and q are known natural numbers. For non-stationary values of d_0 ($d_0 \geq 0.5$), we assume that the process y_t begins[1] at time $t = 1$, that is,

$$y_t = \Delta^{-m_0} x_t(m_0),\ t > 0 \text{ and } = 0 \text{ if } t \leq 0,$$

where

$$\Phi_0(L)\Delta^{\varphi_0}(x_t(m_0) - \mu_0) = \Theta_0(L)\varepsilon_t,\ t = 0, \pm 1, ... \tag{3}$$

In the previous definition, the memory parameter, d_0, is composed as the sum of an integer and a fractional part such that $d_0 = m_0 + \varphi_0$. The integer $m_0 = \lfloor d_0 + 1/2 \rfloor$, where $\lfloor . \rfloor$ denotes integer part, is the number of times that y_t must be differenced to achieve stationarity (therefore $m_0 \geq 0$). The parameter φ_0, the fractional part, lies in the interval $(-0.75, 0.5)$, in such a way that, for a given d_0, $\varphi_0 = d_0 - \lfloor d_0 + 1/2 \rfloor$. Once the process y_t is differenced m_0 times, the differenced process is a stationary fractionally integrated process with an integration order equal to φ_0. For $m_0 = 0$, μ_0 is the expected value of the stationary process y_t and for $m_0 \geq 1$, $\mu_0 \neq 0$ implies a deterministic polynomial trend.

[1] See Marinucci and Robinson (1999) for the different definitions of non-stationary ARFIMA processes and their asymptotic implications.

To derive the new estimator we need to define the residuals of the process. For that purpose, we adopt Beran's (1995) definition of residuals and provide two alternative expressions according to whether the mean, μ_0, is known or unknown.

2.1. Residuals when μ_0 is known and equal to zero.

Let $\psi = \left(\phi_1, ..., \phi_p, \theta_1, ...\theta_q\right)' \in \Re^{p+q}$ be the vector containing the autoregressive and moving average parameters and $\lambda = \left(d, \psi'\right)' \in \Re^{p+q+1}$. Also let Λ be a compact set containing all possible parameter values λ that verify the conditions above and $\lambda_0 = \left(d_0, \psi_0'\right)'$ be an interior point of Λ representing the true parameter values. The (infinite)[2] autoregressive representation of y_t is given by

$$\sum_{j=0}^{\infty} \alpha_j(\lambda_0)\, x_{t-j}(m_0) = \varepsilon_t, \tag{4}$$

where $x_t(m_0) = \Delta^{m_0} y_t$ and $\{\alpha_j(\lambda_0)\}_{j=0}^{\infty}$ are the coefficients associated to the expansion of $\Phi_0(L)\,\Theta_0(L)^{-1}\,\Delta^{\varphi_0}$ in powers of L. Given the observations $y_1, ..., y_T$, the innovations ε_t cannot be computed directly, since an infinite sample would be needed. Nevertheless, they may be estimated by,[3]

$$e_t(\lambda) = \sum_{j=0}^{t-m-1} \alpha_j(\lambda)\, x(m)_{t-j}, \qquad t = m+1, ..., T. \tag{5}$$

[2] This expansion is valid for all $d_0 > -1$. For values of $d_0 > -0.5$, this is a well known result due to Hosking (1981). When $d_0 \in (-1, -0.5)$, Odaki (1993, Theorem 2) shows that although the coefficients $\pi_j(d_0)$ are not square summable and, consequently, the same applies to the coefficients $\alpha_j(\lambda_0)$, the process is still invertible and therefore the autoregressive inversion is well defined.

[3] Notice that for a given d, m and φ are uniquely determined as $\varphi = d - \lfloor d + 1/2 \rfloor$ and $m = d - \varphi$.

2.2. Residuals when μ_0 is unknown.

When μ_0 is unknown the residuals defined above need to be adjusted. Again, following Beran (1995), we consider,

$$\bar{x}(m) = \frac{1}{T-m} \sum_{t=m+1}^{T} x_t(m). \tag{6}$$

Notice that since $x_t(m_0)$ is stationary and ergodic, the sample mean $\bar{x}(m_0)$ is a consistent estimator of μ_0. Therefore, adjusted residuals can be defined as:

$$e_t(\lambda) = \sum_{j=0}^{t-m-1} \alpha_j(\lambda)\left(x_{t-j}(m) - \bar{x}(m)\right), \qquad t = m+1, ..., T. \tag{7}$$

where $\{\alpha_j(\lambda)\}_{j=0}^{\infty}$ are the coefficients associated to the expansion of $\Phi(L)\Theta(L)^{-1}\Delta^{\varphi}$ in powers of L.

3. Generalized Minimum Distance Estimation of ARFIMA processes.

Minimum Distance (MD) is a classical estimation approach in the econometric literature. This technique encompasses other very popular procedures such as Generalized Method of Moments (GMM), Non-Linear Least Squares (NLS) or Maximum Likelihood (ML) among others. In a general framework, this technique would work as follows: If $\lambda_0 \in \Lambda$ is the vector of parameters of interest and y_t is the available data, MD estimation provides a class of estimators that minimize the following criterion function,

$$V_T(\lambda) = \hat{g}_T(\lambda)' \hat{W} \hat{g}_T(\lambda), \tag{8}$$

where $\hat{g}_T(\lambda)$ is a function of the data, y_t , and the parameters of interest, λ, that verifies that $\hat{g}_T(\lambda_0) \xrightarrow{p} 0$; $\hat{W}$ is a positive definite weighting matrix that defines the distance. Under the standard regularity conditions, it can be proved that the resulting estimators are $\sqrt{T}$-consistent and asymptotically normally distributed (see, for instance, Newey and McFaden, 1994). Different choices of the function $\hat{g}_T(\lambda)$ will generate different estimators. For instance, if $\hat{g}_T(\lambda) = T^{-1}\sum_{t=1}^{T} g(y_t, \lambda)$, where $E(g(y_t, \lambda_0)) = 0$, the minimization of the criterion function in (8) would provide a GMM estimator. In particular, the function

$\hat{g}_T(\lambda)$ employed in this paper can be interpreted as the difference between the sample and the population autocorrelations of the residuals defined in (5) or in (7). As it will be proved below, $\hat{g}_T(\lambda)$ defined in this way fulfils the above mentioned requirements. With respect to the choice of $\hat{W}$, it is a well-known result that if $var\left(\sqrt{T}\hat{g}_T(\lambda_0)\right) \xrightarrow{p} \Omega$, then the efficient weighting matrix, W_e, is given by $W_e = \Omega^{-1}$, since in this case the asymptotic variance-covariance matrix of $\hat{\lambda}$ simplifies to $\left(J'_{\lambda_0}\Omega^{-1}J_{\lambda_0}\right)^{-1}$, where J_{λ_0} is the limit of the Jacobian matrix of $\hat{g}_T$ (see Newey and McFaden, 1994). Therefore, this would be in general a good choice. Some examples of $\hat{W}$ corresponding to particular situations will be provided below. As it becomes clear from the discussion above, one of the main advantages of MD relative to ML estimation is that the former do not require to assume a particular distribution of the innovation sequence at any stage.

Several parametric and semiparametric MD techniques can be found in the literature of fractionally integrated (FI) processes. Robinson (1994a) proposed a semiparametric time domain procedure that exploits the fact that autocovariances in FI models are proportional to k^{2d-1} for large k. Robinson´s estimator minimizes the expression: $\sum_{k=n}^{p+n}\left(\hat{\gamma}_k - ck^{2d-1}\right)^2$. Hall et al. (1997) have analyzed the rates of convergence of this estimator but its distributional properties remain to be determined. Of particular relevance to this paper is the MD estimator proposed by Tieslau, Schmidt and Baillie (1996), henceforth TSB. They introduced a parametric time domain MD estimator that minimizes a distance between the estimated and the theoretical autocorrelations of an ARFIMA (p, d, q) process:

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda}\left(\hat{\rho}_{ky} - \rho_{ky}(\lambda)\right)'\hat{W}\left(\hat{\rho}_{ky} - \rho_{ky}(\lambda)\right), \tag{9}$$

where $\hat{\rho}_{ky}$ is the sample autocorrelation function of the (stationary) process y_t up to lag k (for a fixed value of k), $\rho_{ky}(\lambda)$ is the theoretical autocorrelation of the corresponding ARFIMA(p, d, q) process up to the same lag and $\hat{W}$ is a positive definite weighting matrix. The asymptotic optimal weighting matrix is $W_e = C^{-1}$, where C is the asymptotic variance-covariance matrix of $\sqrt{T}\hat{\rho}_{ky}$. Then, a suitable choice for $\hat{W}$ in this context would be a consistent estimator of C^{-1}. Although theoretically very appealing, there remain significant

problems with this procedure. First, it is restricted to stationary series since it requires the existence of autocorrelations. And second, it is $\sqrt{T}-$consistent and asymptotically normal only for $d_0 < 0.25$, due to the non-standard behavior of sample autocorrelations of ARFIMA processes outside this range. Chung and Schmidt (1995) have introduced a modification (Adjusted Minimum Distance Estimator) to the previous estimator. They have demonstrated, by applying the results on autocorrelations of Hosking (1996), that it is possible to obtain a $\sqrt{T}$-consistent and asymptotically normally distributed estimator of d in the whole invertible and stationary range, $-0.5 < d < 0.5$, if some functions of the autocorrelations are employed in the criterion function. Yet, it is only valid in the invertible and stationary range of values of d and it is computationally almost as demanding as exact ML (see Sowell, 1992) since it requires the computation of the autocorrelations as functions of the unknown parameters. Along the same lines, Wright (1999) has proposed an estimator for the fractionally integrated stochastic volatility model and has proved that it is $\sqrt{T}$ -consistent and asymptotically normally distributed only when $d < 0.25$. Galbraith and Zinde-Walsh (1997) have presented a parametric time-domain estimator based on an autoregressive approximation. This estimator can be applied to nonstationary series, since the existence of autocorrelations is not required, but its consistency has not been proved yet in this general framework. [4]

Let us now describe the GMD estimator. Consider the sample ith-autocorrelation associated with the (truncated) residuals defined in (5) or in (7) given by:

$$\hat{\rho}_{e(\lambda)}(i) = \frac{\sum_{t=1}^{T-i} e_t(\lambda) e_{t+i}(\lambda)}{\sum_{t=1}^{T} e_t(\lambda)^2}. \tag{10}$$

Let $\hat{\rho}_{ke(\lambda)}$ be the vector that contains the first k autocorrelations of the residuals, that is,

$$\hat{\rho}_{ke(\lambda)} = \left(\hat{\rho}_{e(\lambda)}(1), \quad ... \quad , \hat{\rho}_{e(\lambda)}(k) \right)'. \tag{11}$$

When evaluated at $\lambda = \lambda_0$, the autocorrelation vector (10) coincides with the sample autocorrelation of the true innovations, ε_t, plus a term that is $o_p\left(T^{-1/2}\right)$ for any k $(< \infty)$.

[4] Other interesting references on the parametric estimation of possibly non-stationary ARFIMA processes are Beran (1995), Tanaka (1999), Velasco and Robinson (2000) and Ling and Li (1997).

This implies that the asymptotic distribution of the sample autocorrelations associated to $e_t(\lambda_0)$ or ε_t is the same. The following theorem formally states this result.

Theorem 1 *Consider the vector defined in* (11) *evaluated at* $\lambda = \lambda_0 \in \Lambda$. *Under the assumptions of Section 2, then,*

$$\sqrt{T}\left(\hat{\rho}_{ke(\lambda_0)} - \hat{\rho}_{k\varepsilon}\right) = o_p(1). \tag{12}$$

where $\hat{\rho}_{k\varepsilon} = \left(\hat{\rho}_{\varepsilon}(1), \quad ... \quad , \hat{\rho}_{\varepsilon}(k) \right)'$ *is the vector that contains the sample first* k *autocorrelations associated to* ε_t *and* k *is fixed.*

The result above together with the Slutsky's Theorem ensure that the asymptotic distribution of $\hat{\rho}_{ke(\lambda_0)}$ and $\hat{\rho}_{k\varepsilon}$ is the same. Hence,

$$\sqrt{T}\hat{\rho}_{ke(\lambda_0)} \xrightarrow{w} N(0, I_k), \text{ for } k = 1, ..., K$$

where I_k is the identity matrix of order k, for a finite value of k . Appendix 2 includes the results of some Monte Carlo experiments that investigate the small sample behavior of this approximation.

Following the argument in TSB (1996), we consider the minimization of a distance between the estimated and theoretical autocorrelations but, in place of the original series, the correlations of the above defined residuals are considered. Since the asymptotic variance of $\sqrt{T}\hat{\rho}_{ke(\lambda_0)}$ is given by I_k, the identity matrix of order k, (Theorem 1), it follows (see (8)) that the efficient weighting matrix is I_k. Moreover, since $\hat{\rho}_{ke(\lambda_0)} \xrightarrow{p} 0$, the MD criterion function, V_{ke}, becomes

$$V_{ke}(\lambda, y) = \hat{\rho}'_{ke(\lambda)}\hat{\rho}_{ke(\lambda)} = \sum_{i=1}^{k} \hat{\rho}_{e(\lambda)}(i)^2, \tag{13}$$

and the GMD estimator $\hat{\lambda}_k$ is defined as:

$$\hat{\lambda}_k = \arg\min_{\lambda \in \Lambda} V_{ke}(\lambda, y). \tag{14}$$

Notice that, since all the sample autocorrelations converge to zero in probability and the efficient weighting matrix is the identity matrix, the criterion function simplifies notably

with respect to that defined in (9), corresponding to the TSB approach. Also, it just requires the existence of the autocorrelations of the residuals but not those of the original series and therefore it can also be applied to non-stationary series, in contrast to the TSB estimator.

So far we have followed previous approaches by considering that the number of autocorrelations included in the criterion function, k, is a fixed number. A further novelty of this paper consists of allowing k to grow to infinity with the sample size T. As it will be seen in Theorems 3 and 4, estimators computed with a fixed k $(> p+q+1)$[5] are consistent and asymptotically normal but they are not efficient. Considering an increasing k allows to obtain an asymptotically efficient estimator (see Theorem 5).

The following theorem extends the results of Theorem 1 by showing that the sum of (weighted) correlations also converges to a Normal distribution when the number of correlations is allowed to go to infinity.

Theorem 2 *Let $\hat{\rho}_{ke(\lambda_0)}$ be the $k \times 1$ vector defined in (11) evaluated at $\lambda = \lambda_0$ and let $\hat{A}_k$ be a $k \times h$ stochastic matrix that verifies that $\hat{A}_k \xrightarrow{p} A_k$ for each fixed k, such that $\sup_{1 \leq i \leq k} (\hat{a}_{ij} - a_{ij}) \xrightarrow{p} 0$ for all $j = 1, ...h$, where $\hat{a}_{ij}$ and a_{ij} are the (i,j) elements of the matrices $\hat{A}_k$ and A_k, respectively. Let k=k(T) be a function of T such that $k(T) \underset{T \to \infty}{\to} \infty$ and $k(T)/T \underset{T \to \infty}{\to} 0$. Finally, let $\Omega = \lim_{k \to \infty} A_k' A_k$ be a positive definite $h \times h$ matrix of real numbers. Then, under the assumptions of this section,*

$$\sqrt{T} \hat{A}'_{k(T)} \hat{\rho}_{k(T)e(\lambda_0)} \xrightarrow{w} N_h(0, \Omega). \tag{15}$$

Later in the paper, this theorem will be used to show the asymptotic normality of the estimator when both k and T tend to infinity. For that purpose, $\hat{A}_k$ and A_k will be replaced by the Jacobian matrix of $\hat{\rho}_{ke(\lambda)}$, that is, the $k \times (p+q+1)$ matrix of partial derivatives of $\hat{\rho}_{ke(\lambda)}$ with respect to λ, and its corresponding probability limit, respectively.

Let us now describe the asymptotic properties of the new estimator. Theorems 3 and 4 state its consistency and asymptotic distribution, respectively, when $T \to \infty$ and k,

[5] This restriction on k should be imposed to achieve identifiability of the parameters.

the number of correlations included in the criterion function, is a fixed number. It turns out that the new estimator is consistent and asymptotically normal. Nevertheless, the asymptotic variance depends negatively on the number of autocorrelations, k, and therefore, the estimator is not efficient. Theorem 5 extends the previous results to the case where k is a function of the sample size that also goes to infinity with T. It is shown that by considering a growing k, the efficiency bound is reached and therefore, the estimator is efficient. As in Beran (1995), Tanaka (1999) and Phillips (1999), and in contrast to Velasco and Robinson (2000), no tapering is required to achieve these results.

Theorem 3 *Let y_t be an ARFIMA(p, d_0, q) process under the hypotheses of Section 2. Also let λ_0 be an interior point of the compact set Λ. Then, as T tends to infinity, it holds that*

$$\hat{\lambda}_k \xrightarrow{p} \lambda_0,$$

where $\hat{\lambda}_k$ is the GMD estimator defined in (14) *and k is a fixed number such that $k \geq p + q + 1$.*

Theorem 4 *Under the hypotheses of the previous theorem, it holds that:*

$$\sqrt{T}(\hat{\lambda}_k - \lambda_0) \xrightarrow{w} N\left(0, \Xi_k^{-1}\right), \tag{16}$$

(for definition of Ξ_k^{-1}, see Appendix 1).

The following theorem presents analogous results when the number of correlations to be included in the criterion function is not a fixed number but it grows with the sample size. By allowing for an increasing k, efficiency is achieved.

Theorem 5 *Let y_t be an ARFIMA(p, d_0, q) process under the hypotheses of Section 2 and let λ_0 be an interior point of the compact set Λ. Let $\hat{\lambda}$ be the GMD estimator defined in* (14) *where the number of correlations included in the criterion function, k, is a function of T such that $\lim_{T\to\infty} k(T) = \infty$ and $\lim_{T\to\infty} k(T)/T = 0$. Then, as T tends to infinity, it holds that*

$$\hat{\lambda} \xrightarrow{p} \lambda_0,$$

and

$$\sqrt{T}(\hat{\lambda} - \lambda_0) \xrightarrow{w} N\left(0, \Xi^{-1}\right), \tag{17}$$

where $\Xi = \lim_{k\to\infty} \Xi_k$ *is the Fisher information matrix (for definition, see Appendix 1).*

The matrix Ξ is identical to the Fisher information matrix for ARMA processes except for the first row and column. Then, estimation of the remaining columns and rows is identical to that of ARMA processes (see, for instance, Brockwell and Davis, (1993)). See also Tanaka (1999) for a description of the estimation of the whole variance-covariance matrix.

It is interesting to recall that many estimation approaches developed for ARFIMA processes are only valid in the invertible and stationary range of values of d, $(-0.5 < d < 0.5)$. In this paper we have enlarged this interval both to the left (up to -0.75) and to the right (for any value of $d < \nabla_2 < \infty$). Other parametric estimators also share this upper bound, such as those presented in Beran (1995) and Velasco and Robinson (2000), in the time and in the frequency domain respectively. Nevertheless, the GMD estimator has some advantages over these approaches. Beran's proof of consistency encounters some problems[6] as Velasco and Robinson (2000) point out that, by adopting a different strategy, it is possible to circumvent in this paper. Velasco and Robinson (2000) show that the Whittle estimate is still appropriate for non-stationary processes as long as enough tapering is applied to the data. However the introduction of tapering increases the variance of the estimators and therefore induces an efficiency loss.

Finally, notice that the estimation approaches that just consider the invertible and stationary range of values of d ($d \in (-0.5, 0.5)$) should first determine m_0 in an exploratory way. In the case where the value of m_0 is correctly guessed, an efficient estimator of this type and the estimator proposed in this paper will share the same asymptotic variance-covariance matrix and therefore they will be (asymptotically) equivalent. But when the assumption about the value of m_0 is wrong, the former family of estimators will become inconsistent while this problem will not affect the GMD estimator. For this reason it is safer

[6] Nevertheless, the reported simulations support Beran's conclusions that this estimator is $\sqrt{T}$-consistent, asymptotic normally distributed and efficient for all $d > -0.5$.

to use a method that covers the whole range of values of d and there is no loss of efficiency by doing so if the GMD estimator is employed.

4. Behavior of the GMD estimator in finite samples.

It follows from Theorem 5 that, asymptotically, the larger the value of k chosen to compute $\hat{\lambda}_k$, the better. But in general, this result may not to be true in finite samples. In this section, a Monte Carlo study is conducted to investigate the small-sample performance of the MD estimator defined in (14) and to determine how to select k.

Processes of the form $\Delta^{\varphi_0}(\Delta^{m_0} y_t - \mu_0) = u_t$ have been generated, with four different specifications for u_t, namely $u_t = \varepsilon_t$, $u_t = \phi_1 u_{t-1} + \varepsilon_t$, $u_t = (1 + \theta_1 L)\varepsilon_t$, and $u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \varepsilon_t$, $\varepsilon_t \sim NID(0,1)$ in all cases. In particular, μ_0 is set equal to zero but the estimation procedure is carried out both considering that its value is known (and equal to zero) or unknown.

Before estimating the above-mentioned models, it is necessary to select the number of correlations (k) to be included in the criterion function V_{ke}. In finite samples, the suitable choice of k depends on the sample size T, on the number $(p+q+1)$ of parameters and on the values of these parameters. Asymptotic theory does not help much with respect to the right choice of k and, therefore, this is a question that should be addressed via Monte Carlo simulation. Table 1 presents the bias and the square root of the mean square error ($\sqrt{MSE}$) of the GMD estimator in the case where $u_t = \varepsilon_t$, for two different sample sizes, $T = 100$ and $T = 400$. The number of replications was 1000. Different values of k were used, namely, the closest integer to the quantities: $T^{1/4}$, $T^{1/3}$, $T^{1/2}$ (more precisely $k = 3, 5, 10$ and $k = 4, 7, 20$ for $T = 100$, 400 respectively). Table 2 presents analogous results for the case where μ_0 is unknown and has to be estimated. Tables 3, 4 and 5 display the figures obtained from similar experiments for the case where u_t is an $AR(1)$, $MA(1)$ or an $AR(2)$ process with parameters $\phi_1 = 0.6$, $\theta_1 = 0.5$ or $\phi_1 = 0.65$ and $\phi_2 = -0.6$ respectively. The value of the parameters have been chosen in order to facilitate comparison with the ones used in previous studies.

Tables 1 to 5 describe the performance in finite samples of the proposed method. Tables

1 and 2 show that the GMD estimator is very robust to different values of k. Moderate values of k provide the best results. For $T = 100$, the $\sqrt{MSE}$ is in general smaller for those estimates computed with $k = T^{1/4}$, although for $T = 400$ the $\sqrt{MSE}$ is very similar and even smaller for those computed with $T^{1/3}$. With respect to efficiency, the asymptotic (efficient) standard deviation of $\hat{d}$ in the ARFIMA$(0, d, 0)$ model is $\pi^{-1}\sqrt{(6/T)}$, which equals 0.078 and 0.039 for $T = 100$ and 400, respectively. The reported $\sqrt{MSE}s$ in Table 1 approximate quite well these values. It is also remarkable the small bias of the estimates, even for very large values of d. The bias is usually negative which suggests that the GMD method slightly underestimates the memory parameter. Also notice that the figures for the case where μ_0 is unknown and has to be estimated do not differ significantly from the case where it is known. For the AR and MA cases, different values of k were used and similar results were obtained. Again, the estimator performs better in terms of MSE when moderate values of k are used in the criterion function. The estimates in Table 3 to 5 have been calculated with $k = T^{1/4}$.[7] The asymptotically efficient variance-covariance matrix ($Asymp.\ Var$) for the ARFIMA$(1, d, 0)$ process is given by,

$$Asymp.\ Var \begin{pmatrix} \hat{d} \\ \hat{\phi}_1 \end{pmatrix} = \frac{1}{T} \begin{pmatrix} \frac{\pi^2}{6} & -\frac{1}{\phi_1}\log(1-\phi_1) \\ -\frac{1}{\phi_1}\log(1-\phi_1) & \frac{1}{1-\phi_1^2} \end{pmatrix}^{-1}, \qquad (18)$$

which delivers, for a value of $\phi_1 = 0.6$, asymptotic standard deviations equal to 0.256 and 0.262 for $T = 100$, and 0.128 and 0.131 for T=400, corresponding to $\hat{d}$ and $\hat{\phi}_1$ respectively. As for the ARFIMA$(0, d, 1)$ case, asymptotic standard deviations can be computed from an analogous expression and, for $\theta_1 = 0.5$, they are equal to 0.222 and 0.246 for $T = 100$ and 0.1108, and 0.1231 for $T = 400$ corresponding also to $\hat{d}$ and $\hat{\theta}_1$ respectively. It can be seen again that the reported $\sqrt{MSE}s$ are a reasonable approximation to these values.

Table 6 compares the $\sqrt{MSE}$ of different estimators of d for the ARFIMA$(0, d, 0)$ case. More specifically we consider the Whittle estimator (with Zhurbenko taper of order 2) proposed by Velasco and Robinson (VR), the ML estimators proposed by Sowell (SOW)

[7] Other values of k have been also tried and the figures are available upon request.

and Beran (BER) and the minimum distance estimators by Tieslau et al. (TSB) and by Galbraith and Zinde-Walsh (GZW). The DGP used in this experiment was a fractional white noise with known mean equal to zero. The missing values in Table 6 stem from methods that are not defined for the whole range of values of d. It can be observed that the GMD estimator behaves similarly to the ML estimators and better than the remaining ones. It is also remarkable the good performance of the GMD estimator in the range of values of d in which other estimators are not defined.

Table 1. Estimation of d for the $ARFIMA(0, d, 0)$ case

μ_0 **known** ($\mu_0 = 0$)										
d_0		-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$										
	$k = T^{1/4}$	-0.018	-0.032	-0.021	-0.017	-0.031	-0.019	-0.019	-0.026	-0.021
bias $\hat{d}$	$k = T^{1/3}$	-0.020	-0.031	-0.022	-0.029	-0.035	-0.021	-0.21	-0.028	-0.023
	$k = T^{1/2}$	-0.022	-0.032	-0.024	-0.032	-0.036	-0.023	-0.032	-0.030	-0.025
	$k = T^{1/4}$	0.101	0.105	0.094	0. 101	0.0970	0.093	0.093	0.104	0.098
$\sqrt{MSE}$ $\hat{d}$	$k = T^{1/3}$	0.111	0.110	0.111	0.109	0.114	0.103	0.103	0.118	0.102
	$k = T^{1/2}$	0. 116	0.117	0.114	0.113	0. 122	0.116	0. 118	0.122	0.121
$T = 400$										
	$k = T^{1/4}$	-0.003	-0.007	-0.008	-0.010	-0.008	0.008	0.005	0.006	0.008
bias $\hat{d}$	$k = T^{1/3}$	-0.003	-0.008	-0.010	-0.012	0.008	0.010	0.006	0.006	0.002
	$k = T^{1/2}$	-0.003	-0.008	-0.010	-0.013	0.010	0.010	0.007	0.008	0.009
	$k = T^{1/4}$	0.045	0.042	0.045	0.041	0.042	0.046	0.044	0.044	0.046
$\sqrt{MSE}$ $\hat{d}$	$k = T^{1/3}$	0.044	0.041	0.043	0.041	0.042	0.046	0.043	0.044	0.044
	$k = T^{1/2}$	0.047	0.042	0.045	0.042	0.044	0.048	0.044	0.045	0.046

Table 2. Estimation of d for the $ARFIMA(0, d, 0)$ case

		μ_0 *unknown* ($\mu_0 = 0$)								
d_0		-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$										
	$k = T^{1/4}$	-0.005	-0.030	0.002	-0.003	-0.026	-0.022	-0.024	-0.027	-0.020
bias $\hat{d}$	$k = T^{1/3}$	-0.007	-0.036	-0.002	-0.031	-0.027	-0.025	-0.026	-0.030	-0.022
	$k = T^{1/2}$	-0.009	-0.034	-0.027	-0.031	-0.030	-0.027	-0.029	-0.029	-0.025
	$k = T^{1/4}$	0.101	0.107	0.097	0.103	0.104	0.100	0.102	0.104	0.099
$\sqrt{MSE}\ \hat{d}$	$k = T^{1/3}$	0.113	0.107	0.106	0.109	0.110	0.114	0.104	0.107	0.105
	$k = T^{1/2}$	0.117	0.109	0.111	0.112	0.123	0.117	0.117	0.117	0.109
$T = 400$										
	$k = T^{1/4}$	0.008	-0.007	0.001	-0.05	-0.008	-0.010	-0.004	-0.009	-0.008
bias $\hat{d}$	$k = T^{1/3}$	0.009	-0.009	0.003	-0.007	-0.007	-0.010	-0.009	-0.010	-0.010
	$k = T^{1/2}$	0.022	-0.009	-0.001	-0.007	-0.008	-0.011	-0.009	-0.012	-0.016
	$k = T^{1/4}$	0.046	0.042	0.046	0.044	0.044	0.046	0.044	0.042	0.043
$\sqrt{MSE}\ \hat{d}$	$k = T^{1/3}$	0.047	0.042	0.049	0.043	0.444	0.049	0.048	0.040	0.043
	$k = T^{1/2}$	0.052	0.043	0.049	0.044	0.049	0.050	0.049	0.045	0.045

Table 3. Estimation of $\lambda = (d, \boldsymbol{\phi}_1)'$.

DGP: $ARFIMA(1,d,0)$, $\phi_1 = 0.6$, $k = T^{1/4}$									
μ_0 **known** $(\mu_0 = 0)$									
d_0	-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$									
bias $\hat{\mathrm{d}}$	-0.030	-0.018	-0.044	0.025	-0.046	-0.053	-0.053	-0.034	-0.055
$\sqrt{MSE}$ $\hat{\mathrm{d}}$	0 212	0.225	0.240	0.230	0.253	0.234	0.237	0.245	0.255
bias $\hat{\phi}_1$	-0.013	-0.023	-0.062	0.000	0.006	0.007	0.011	0.021	0.008
$\sqrt{MSE}$ $\hat{\phi}_1$	0.219	0.249	0.253	0.203	0.203	0.202	0.193	0.215	0.209
$T = 400$									
bias $\hat{\mathrm{d}}$	-0.028	0.028	0.012	0.023	0.024	0.053	0.027	0.022	0.054
$\sqrt{MSE}$ $\hat{\mathrm{d}}$	0.150	0.140	0.136	0.158	0.142	0.164	0.150	0.155	0.162
bias $\hat{\phi}_1$	0.001	0.002	-0.010	0.000	0.003	0.029	0.010	0.007	0.018
$\sqrt{MSE}$ $\hat{\phi}_1$	0.142	0.135	0.135	0.148	131 03	0.145	0.146	0.141	0.143
μ_0 **unknown** $(\mu_0 = 0)$.									
d_0	-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$									
bias $\hat{\mathrm{d}}$	-0.022	-0.021	0.000	-0.011	-0.045	-0.021	-0.032	-0.023	-0.022
$\sqrt{MSE}$ $\hat{\mathrm{d}}$	0.219	0.221	0.181	0.219	0.221	0.221	0.210	0.233	0.222
bias $\hat{\phi}_1$	-0.024	-0.043	-0.041	-0.025	0.003	0.000	-0.010	0.010	0.006
$\sqrt{MSE}$ $\hat{\phi}_1$	0.211	0.197	0.182	0.182	0.216	0.190	0.210	0.215	0.227
$T = 400$									
bias $\hat{d}$	-0.004	-0.031	-0.018	-0.024	-0.023	-0.013	-0.015	-0.010	-0.017
$\sqrt{MSE}$ $\hat{d}$	0.146	0.141	0.131	0.143	0.140	0.143	0.143	0.144	0.147
bias $\hat{\phi}_1$	0.018	0.004	0.013	0.001	0.003	0.026	0.013	-0.016	0.009
$\sqrt{MSE}$ $\hat{\phi}_1$	0.143	0.136	0.127	0.147	0.132	0.139	0.138	0.140	0.141

Table 4. Estimation of $\lambda = (d, \boldsymbol{\theta}_1)'$.

DGP: **ARFIMA**$(\mathbf{0}, \mathbf{d}, \mathbf{1})$, $\boldsymbol{\theta}_1$= **0.5**. $k = T^{1/4}$

μ_0 **known** $(\mu_0 = 0)$

d_0	-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$									
bias $\hat{d}$	-0.012	-0.016	-0.010	-0.013	-0.018	-0.012	0.0206	0.004	0.025
$\sqrt{MSE}$ $\hat{d}$	0.148	0.131	0.148	0.147	0.138	0.164	0.244	0.213	0.260
bias $\hat{\theta}_1$	-0.000	0.002	-0.003	-0.003	0.008	-0.003	-0.084	-0.146	-0.339
$\sqrt{MSE}$ $\hat{\theta}_1$	0.146	0.127	0.138	0.141	0.131	0.141	0.2158	0.255	0.435
$T = 400$									
bias $\hat{d}$	-0.000	-0.007	-0.007	-0.005	-0.0082	-0.012	-0.0037	-0.009	-0.012
$\sqrt{MSE}$ $\hat{d}$	0.056	0.058	0.057	0.0574	0.056	0.0552	0.0567	.0591	0.077
bias $\hat{\theta}_1$	0.00	0.005	0.0029	-0.003	0.0045	.0036	-0.025	-0.04700	-0.1834
$\sqrt{MSE}$ $\hat{\boldsymbol{\theta}}_1$	0.056	0.0542	0.0590	0.0582	00.0567	0.0541	0.0699	0.0974	0.2453

μ_0 **unknown** $(\mu_0 = 0)$.

d_0	-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$									
bias $\hat{d}$	-0.019	-0.021	-0.008	-0.016	-0.021	-0.018	-0.023	-0.022	-0.0126
$\sqrt{MSE}$ $\hat{d}$	0.135	0.129	0.147	0.145	0.140	0.148	0.139	0 .146	0.146
bias $\hat{\theta}_1$	0.002	0.007	-0.005	-0.000	0.012	0.002	0.008	0.007	-0.006
$\sqrt{MSE}$ $\hat{\theta}_1$	0.129	0.125	0.135	0.136	0.134	0.135	0.122	0.139	0.142
$T = 400$									
bias $\hat{d}$	-0.004	-0.031	-0.018	-0.024	-0.023	-0.013	-0.015	-0.010	-0.017
$\sqrt{MSE}$ $\hat{d}$	0.146	0.141	0.131	0.143	0.140	0.143	0.143	0.144	0.147
bias $\hat{\theta}_1$	0.018	0.004	0.013	0.001	0.003	0.026	0.013	-0.016	0.009
$\sqrt{MSE}$ $\hat{\theta}_1$	0.143	0.136	0.127	0.147	0.132	0.139	0.138	0.140	0.141

Table 5. Estimation of $\lambda = (d, \phi_1, \phi_2)'$.

DGP: **ARFIMA(2, d, 0)**, ϕ_1= **0.6**, ϕ_2= **−0.65**;. $k = T^{1/4}$

μ_0 **known** $(\mu_0 = 0)$									
d_0	-0.7	-0.3	0.4	0.8	1.0	1.4	1.8	2.0	2.4
$T = 100$									
bias $\hat{d}$	-0.003	-0.031	-0.030	-0.021	-0.022	-0.064	-0.113	-0.168	-0.112
$\sqrt{MSE}$ $\hat{d}$	0.224	0.228	0.230	0.229	0.223	0.225	0.244	0.281	0.282
bias $\hat{\phi}_1$	0.027	0.074	0.052	0.031	0.041	0.058	0.043	0.049	0.051
$\sqrt{MSE}$ $\hat{\phi}_1$	0.206	0.219	0.232	0.216	0.226	0.230	0.261	0.302	0.306
bias $\hat{\phi}_2$	-0.059	-0.050	-0.083	-0.069	-0.072	-0.067	-0.053	-0.031	-0.038
$\sqrt{MSE}$ $\hat{\phi}_2$	0.150	0.171	0.174	0.159	0.165	0.165	0.202	0.371	0.382
$T = 400$									
bias (d̂)	0.004	-0.011	-0.010	-0.006	-0.012	-0.015	-0.032	-0.051	-0.112
$\sqrt{MSE}$ $\hat{d}$	0.067	0.063	0.061	0.063	0.064	0.064	0.074	0.087	0.149
bias $\hat{\phi}_1$	-0.003	0.003	0.001	0.001	0.003	0.002	-0.005	-0.019	-0.063
$\sqrt{MSE}$ $\hat{\phi}_1$	0.054	0.054	0.049	0.051	0.053	0.053	0.059	0.067	0.110
bias $\hat{\phi}_2$	0.003	0.003	-0.003	-0.003	-0.003	0.001	0.012	0.024	0.089
$\sqrt{MSE}$ $\hat{\phi}_2$	0.047	0.049	0.048	0.049	0.047	0.048	0.052	0.061	0.126

Table 6. $\sqrt{MSE}$ for various estimation techniques

$T = 100$; DGP: $\Delta^{d_0} y_t = \varepsilon_t$, μ_0 known							
d_0	-0.3	0.2	0.4	0.8	1.0	1.4	1.8
VR	0.211	0.208	0.193	0.172	0.155	0.154	0.160
BER	0.084	0.083	0.089	0.091	0.086	0.092	0.095
TSB	0.189	0.094	0.109	-	-	-	-
GZW	0.082	0.089	0.096	-	-	-	-
SOW	0.089	0.092	0.069	-	-	-	-

5. Residual-based statistics for diagnosing checking.

Although parametric estimates present in general better properties than semiparametric ones, these good properties rely heavily on the correct specification of the parametric model. For this reason, a formal test to check the adequacy of the proposed model is very often carried out. A common way of testing it is by checking the assumption of white noise residuals (see Milhoj (1981) and Chen and Deo (2004) for other approaches that do not require the computation of residuals from the fitted model). Box and Pierce (1970) introduced a goodness-of-fit procedure that tests for significant residual autocorrelations. The Box-Pierce (BP) statistic is defined as

$$Q(k) = T\sum_{i=1}^{k} \hat{\rho}^2_{e(\hat{\lambda})}(i). \tag{19}$$

They proved that, in the context of ARIMA processes, this statistic is asymptotically χ^2-distributed with $k-p-q$ degrees of freedom, for large k. Ljung and Box (1978) introduced a modification to the $Q(k)$ −statistic that improves the approximation to the χ^2_{k-p-q} distribution. This is defined as

$$\tilde{Q}(k) = T(T+2)\sum_{i=1}^{k}(T-k)^{-1}\hat{\rho}^2_{e(\hat{\lambda})}(i).$$

Hong (1996) proposed a generalization of the BP test given by

$$H_T = \left(T\sum_{i=1}^{k(T)} h^2(i/k)\,\hat{\rho}^2_{e(\hat{\lambda})}(i) - C_T(h) \right) / 2D_T(h)^{1/2},$$

where $h(.)$ is a suitable chosen kernel and $k(T)$ verifies that $\lim_{T\to\infty} k(T) = \infty$ and $\lim_{T\to\infty} k(T)/T \to 0$ (for the definition of C_T and D_T see Hong, 1996). When $h(.)$ is the truncated kernel, i.e., $h(z) = 1$ for $|z| \leq 1$ and 0 for $|z| > 1$, it is obtained

$$H^*_T = \left(T\sum_{i=1}^{k(T)} \hat{\rho}^2_{e(\hat{\lambda})}(i) - k(T) \right) / (2k(T))^{1/2}, \tag{20}$$

a generalization of BP's test when $k(T) \to \infty$. Hong (1996) establishes the asymptotic normality of H_T for AR models when $k(T) \to \infty$ and $k(T)/T \to 0$.

The estimation procedure described in Section 3 provides an immediate check on the adequacy of the specified model, since the criterion function defined in (13), evaluated at the estimated values and multiplied by the sample size, coincides with the BP goodness-of-fit statistic in (19). Due to its simplicity, it is worth analyzing the behavior of this statistic. In the context of stationary ARFIMA processes with known mean, the asymptotic distribution of residual autocorrelations have been examined by Li and McLeod (1986). The following theorem is an extension of Li and McLeod's result to the case where the process is allowed to be nonstationary and to have an unknown mean.

Theorem 6 *Let* $\hat{\rho}_{_{ke}(\hat{\lambda})}$ *be the vector containing the autocorrelations up to lag* k *of the residuals, defined in* (5) *or* (7), *such that:*

$$\hat{\rho}_{_{ke}(\hat{\lambda})} = \left(\begin{array}{cccc} \hat{\rho}_{e(\hat{\lambda})}(1), & ... & , \hat{\rho}_{e(\hat{\lambda})}(k) \end{array} \right)' \tag{21}$$

where $\hat{\rho}_{e(\hat{\lambda})}(i)$ *is defined as in* (10). *Then, for any fixed* k, $\sqrt{T}\hat{\rho}_{_{ke}(\hat{\lambda})}$ *is asymptotically normally distributed, with zero mean and variance-covariance matrix given by:*

$$\Sigma = I_k - J_k(\lambda_0)\left(J_k'(\lambda_0) J_k(\lambda_0)\right)^{-1} J_k'(\lambda_0)$$

where $J_k(\lambda_0)$ *is the limit as* T *tends to infinity of the Jacobian matrix of* $\hat{\rho}_{e(\lambda_0)}$.

Applying standard results, it is easily seen that Σ is approximately idempotent with rank $k-p-q-1$, for k large enough. Hence, both $Q(k)$ and $\tilde{Q}(k)$ are approximately χ^2 -distributed with $k-p-q-1$ degrees of freedom for large T.

The previous results imply that the minimum value of the criterion function (13), $V_{ke}\left(\hat{\lambda}\right)$, can be used to test the adequacy of the specified model. Under the null hypothesis of correct specification, the above-mentioned value multiplied by the sample size is distributed approximately $\chi^2_{k-p-q-1}$, for large T, where k is the number of autocorrelations considered in the criterion function. It is also straightforward to compute the Hong's statistic in (20) from this value, just by multiplying it by the sample size, substracting the number of included autocorrelations and dividing by the squared root of $2k(T)$. Although the asymptotic properties of this statistic remain unknown where long memory processes are considered

and their derivation goes beyond the scope of this paper, the following section explores by simulation its finite sample behavior. It also discusses the finite sample properties of both $Q(k)$ and $\tilde{Q}(k)$ statistics.

6. Behavior of the Goodness-of-fit tests in finite samples

To evaluate the performance in terms of size and power of the goodness-of-fit tests examined in Section 5, the following experiments have been carried out. First, processes of the form $\Delta^{d_0} y_t = \varepsilon_t$, $\varepsilon_t \sim NID(0,1)$ were generated for different values of d_0. This parameter was estimated in accordance with the method presented in Section 3 and the $Q(k)$ statistic was computed using the corresponding residuals. The value k was set equal to $3, 4$ and 5 for sample sizes $T = 150$, 400 and 500, respectively (that is, $k \sim T^{1/4}$). Empirical size at the 5% signification level is calculated using the $\chi^2_{k-1,0.95}$ value. Since the Ljung-Box statistic improves the approximation to the χ^2 distribution, it is usually preferred to the Box-Pierce statistic in applications. Therefore, the values of the $\tilde{Q}(k)$ are also computed in order to compare the behavior of both tests. Table 7 reports the empirical size of both the Box-Pierce (BP) and Ljung-Box (LB) tests. In agreement with the findings of Li and McLeod (1986) , the empirical size is close to the nominal size in both tests. Although the approximation to the χ^2_{k-1} distribution is slightly better for the LB test, the difference between both tests decreases as T increases.

With respect to the power of the test, it will obviously depend on how close is the DGP to the null hypothesis. An ARFIMA$(1, d_0, 0)$ is been chosen as the true DGP with a value of the autoregressive parameter ϕ_1 equal to 0.5, for different values of d_0. An ARFIMA$(0, d, 0)$ was estimated instead. Table 8 reports the power of both the Box-Pierce (BP) and Ljung-Box (LB) tests at the 5% nominal signification level. Therefore power has been calculated with the asymptotic critical values and it has not been size-adjusted. Since both tests are in general undersized, better results would have been achieved after size adjustment. It is seen both tests perform quite similarly. For small sample sizes ($T = 150$) the power is similar to that obtained by other methods proposed in the literature (see, for instance, Delgado and Hidalgo, (1999)). It improves considerably when larger samples sizes are considered,

providing good results for sample sizes around 400 or 500.

Table 7. Empirical Size of BP and LB. ($S.L.: 5\%$)

d_0	-0.7	-0.3	0	0.4	0.8	1.4
$T = 150$						
BP	3.5%	3.2%	3.8%	4.0%	3.6%	3.8%
LB	4.5%	4.0%	4.5%	4.1%	4.0%	4.5%
$T = 400$						
BP	6.0%	4.2%	4.8%	4.8%	5.8%	4.5%
LB	6.0%	5.8%	3.5%	5.5%	6.0%	5.0%
$T = 500$						
BP	5.5%	5.0%	5.0%	5.0%	4.5%	4.5%
LB	6.0%	5.0%	5.0%	5.5%	5.0%	4.2%

Table 8. Power of BP and LB. (S.L.: 5%)

DGP: ARFIMA$(1, d, 0)$						
d_0	-0.7	-0.3	0	0.4	0.8	1.4
$T = 150$						
BP	24.8%	24.2%	24.4%	23.0%	23.8%	21.8%
LB	25.3%	24.4%	27.0%	25.2%	25.4%	21.8%
$T = 400$						
BP	66.6%	68.6%	66.9%	64.4%	68.9%	64.5%
LB	67.3%	69.2%	67.9%	65.6%	70.8%	65.3%
$T = 500$						
BP	81.8%	81.2%	79.8%	82.2%	82.4%	82.8%
LB	81.6%	81.6%	80.0%	83.0%	82.2%	83.4%

Tables 9 and 10 report the results of analogous simulations based on Hong's statistic defined in (20). To compute the size, critical values from a $N(0, 1)$ distribution have been employed. Since in this case k is allowed to go to infinity, different values of k have been

employed, namely $k(T) = T^{1/4}$, $T^{1/3}$ and $T^{1/2}$. It is seen that the empirical size is lower that the nominal one but the approximation improves for faster $k(T)$ and larger sample sizes. Size-corrected power is been reported in order to facilitate comparison among different values of k. Large sample sizes, ($T = 400$, 500), are needed to obtain a reasonable power. Also, it can be observed that a slower $k(T)$ provides better power, result which agrees with Hong's (1996) findings.

Table 9. Empirical Size of H_T^*. (S.L.: 5%)

d_0	-0.7	-0.3	0	0.4	0.8	1.4
$T = 150$						
k=$T^{1/4}$	1.2%	1.5%	1.4%	1.1%	1.2%	1.2%
k=$T^{1/3}$	2.2%	2.2%	2.3%	1.8%	1.6%	1.5%
k=$T^{1/2}$	2.5%	2.5%	3.4%	2.8%	2.8%	2.6%
$T = 400$						
k=$T^{1/4}$	1.8%	1.9%	1.5%	2.2%	2.4%	2.1%
k=$T^{1/3}$	2.9%	3.2%	2.1%	2.6%	2.9%	1.5%
k=$T^{1/2}$	5.4%	3.9%	3.1%	3.6%	3.9%	3.2%
$T = 500$						
k=$T^{1/4}$	2.7%	2.1%	2.0%	3.1%	2.9%	2.6%
k=$T^{1/3}$	2.4%	2.3%	2.6%	2.5%	2.6%	3.2%
k=$T^{1/2}$	3.9%	4.2%	3.5%	4.0%	4.5%	5.2%

Table 10. Power of H_T^* (size corrected).

DGP: ARFIMA(1,d_0,0)						
d_0	-0.7	-0.3	0	0.4	0.8	1.4
$T = 150$						
k=$T^{1/4}$	22.3%	16.2%	19.8%	21.2%	20.6%	19.3%
k=$T^{1/3}$	20.7%	14.7%	14.2%	14.2%	16.1%	16.1%
k=$T^{1/2}$	12.1%	10.4%	9.2%	10.1%	13.8%	13.5%
$T = 400$						
k=$T^{1/4}$	54.2%	28.7%	65.4%	54.3%	54.9%	56.5%
k=$T^{1/3}$	55.9%	56.4%	65.3%	56.4%	56.5%	59.8%
k=$T^{1/2}$	30.4%	32.6%	39.3%	35.0%	35.5%	30.6%
$T = 500$						
k=$T^{1/4}$	70.9%	82.1%	76.7%	75.0%	74.1%	76.8%
k=$T^{1/3}$	75.9%	77.7%	68.6%	76.6%	77.1%	69.9%
k=$T^{1/2}$	45.2%	52.8%	45.8%	46.1%	43.2%	39.7%

7. Empirical applications.

In order to illustrate the application of the techniques proposed in this paper, two empirical studies have been carried out. Firstly, to provide a further comparison with previous estimation techniques, the empirical series analyzed by Beran (1995), Velasco and Robinson (2000) and Robinson (1994b) have been considered, namely the chemical process temperature readings (Series C) and the chemical process concentration readings (Series A) from Box and Jenkins (1976). Secondly, the orders of fractional integration of the *GDP* per capita series of several countries have been estimated.

With respect to the first application, our conclusions are in fair agreement with those of the above-mentioned studies. For series C, Box and Jenkins (BJ), fitted an ARIMA$(1,1,0)$ model, with an estimated value of the AR parameter $\phi_1 = 0.8$. The corresponding estimates for an ARFIMA$(1,d,0)$ process according with the GMD method are: $\hat{d} = 1.005$ with a 95%

confidence interval (C.I.) of $[0.7497, 1.2617]$ and $\hat{\phi}_1$ is 0.798 with a 95% C.I. of $[0.563, 0.972]$, in close agreement with BJ's conclusions. Nevertheless, the recognition of the uncertainty on d, increases substantially the standard deviation of the AR parameter. Similar conclusions are reached both in Beran (1995) and Velasco and Robinson (2000).

For series A, BJ fitted an ARIMA$(0,1,1)$, which yields a significative value for the MA parameter equal to -0.7. This large negative value suggests that the model could be overdifferenced. If an ARFIMA$(0,d,1)$ is fitted instead, the GMD estimates are 0.43 and -0.038 with 95% C.I.s of [0.241,0.612] and [-0.27,0.196] for $\hat{d}$ and $\hat{\theta}_1$, respectively. Therefore, $d = 1$ is not included in the C.I., which reinforces the believe that the series in the BJ model is overdifferenced. Since the MA parameter is not significant in this second case, we have also fitted an ARFIMA$(0,d,0)$ to the data. In this case, the estimated value of d drops to 0.401 with a 95% C.I. of [0.292,0.51]. Also in this case, our results follow closely the ones obtained in Velasco and Robinson (2000) and Beran (1995).

In the second application, some of the series of Maddison's (1995) data set have been analyzed. This data set contains the annual GDP per capita series for OECD countries during the period 1870-1994 (125 observations). This data set has also been analyzed by Micchelaci and Zaffaroni (2000) and Dolado et al. (2003) among others. In particular, three countries have been considered, namely Canada, Japan and Germany. Since these series are clearly trended, the estimation has been carried out as if the mean was different from zero and unknown. They have been estimated according to the following parametric methods: the GMD method presented in this paper, the Sowell's exact ML procedure (with the series in first differences when they are nonstationary), the NLS method by Beran (1995) and the Velasco and Robinson (2000) Whittle estimator (with Zhurbenko taper of order 2). Also, several ARFIMA(p,d,q) processes have estimated for values of p, q in the range 0 to 2, and the results reported correspond to the preferred model according to the AIC lag-length criterion. Table 11 reports the value of the estimates and their standard errors (in brackets). As it can be seen from this table, all methods deliver very similar estimates of the parameters. With the exception of Japan, it seems to be the case that Canada and Germany are clear cases where the GDP per capita series are fractionally integrated, with

a value of d around 0.5 for the former and in the range $d \in (0.5, 1)$ for the latter. It is also noticeable that there is a negative relation between the values of d and the values of the remaining parameters across the different methods. Standard deviations are also similar across methods with the exception of the Velasco and Robinson technique which shows slightly higher values due to the tapering employed.

Table 11. Estimation results from various estimation techniques

	Canada	Japan	Germany
Model (AIC)	ARFIMA$(1, d, 0)$	ARFIMA$(0, d, 0)$	ARFIMA$(0, d, 1)$
GMD	$\hat{d} = 0.40; \hat{\phi}_1 = 0.78$ (0.217; 0.174)	$\hat{d} = 1.083$ (0.006)	$\hat{d} = 0.87; \hat{\theta}_1 = 0.49$ (0.195; 0.218)
SOW	$\hat{d} = 0.48; \hat{\phi}_1 = 0.72$ (0.239; 0.213)	$\hat{d} = 1.064$ (0.006)	$\hat{d} = 0.80; \hat{\theta}_1 = 0.51$ (0.201; 0.222)
BER	$\hat{d} = 0.50; \hat{\phi}_1 = 0.67$ (0.242; 0.230)	$\hat{d} = 1.067$ (0.006)	$\hat{d} = 0.78; \hat{\theta}_1 = 0.55$ (0.214; 0.229)
V-R	$\hat{d} = 0.41; \hat{\phi}_1 = 0.64$ (0.345; 0.340)	1.030 (0.101)	$\hat{d} = 0.62; \hat{\theta}_1 = 0.62$ (0.330; 0.341)

8. Conclusions

In this paper we have proposed a new method for estimating the parameters of an ARFIMA(p, d_0, q) process with $d_0 > -0.75$ in the time domain. It covers a very wide range of values of d_0, providing therefore a unified framework for the construction of confidence intervals and tests for the memory parameter. The proposed estimator belongs to the MD class and it is based on the minimization of the squared correlations of the residuals which are obtained after filtering a process through ARFIMA parameters. Its asymptotic properties as well as its finite sample performance are discussed and it is shown that it is $\sqrt{T}$-consistent and asymptotically normally distributed without making strong assumptions on the distribution of the process under study. Furthermore, if the number of autocorrelations is allowed to increase with the sample size, then the estimator is also asymptotically efficient. Monte Carlo experiments show that it is also well-behaved in finite samples and that it compares well to other existing estimators in the literature. Another interesting feature of the estimator is that the criterion function evaluated in the estimate coincides with the Box-Pierce (1970) goodness-of-fit statistic, providing therefore, and immediate tool to

evaluate the adequacy of the model specification. The asymptotic properties of this statistic, as well as the ones of the Ljung-Box (1978) statistic, are discussed and some simulations are provided in order to evaluate their accuracy in finite samples.

Finally, another nice attribute of the proposed estimator is its flexibility to be extended to more general settings. For instance, the estimator can be easily robustified against conditional heterocedasticity, simply by considering the sample autocorrelations of the standardized residuals. Also, following the lines of Wright (1999), it can be adapted to deal with the fractionally integrated stochastic volatility model. Further research should also be undertaken to extend the previous framework to the multivariate case.

References

Amemiya, T. (1985) *Advanced econometrics*, Basil Blackwell.

Baillie, R.T. (1996) "Long memory processes and fractional integration in Economics and Finance". *Journal of Econometrics*, 73, 5-59.

Beran, J. (1995) "Maximum likelihood estimation of the differencing parameter for invertible and short and long memory autoregressive integrated moving average models". *Journal of the Royal Statistical Society*, Series B, 57, No. 4, 659-672.

Bernstein, S. (1927) "Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendentes". *Mathematical Annals*, 97, 1.

Box, G.E.P. and G.M. Jenkins (1976) *Time Series Analysis: Forecasting and Control*. San Francisco, Holden Day.

Box, G.E.P and D. A. Pierce (1970) "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models". *Journal of the American Statistical Association, 6*5, 1509-26.

Brockwell, P. J. and R.A. Davis (1993) *Time series: Theory and methods*, 2nd ed. Springer-Verlag, New York, NY.

Brown, B. M. (1971), "Martingale central limit theorems", *Annals of Mathematical Statistics*, 42, 59-66.

Chen W. W. and R.S. Deo (2004) "A generalized portmanteau Goodness-of-fit test for time series models". *Econometric Theory*, Vol 20 (2), 382-416.

Chung C.F. and P. Schmidt (1995) "The minimum distance estimator for fractionally integrated ARMA processes". Econometrics and Economic Theory Papers, No. 9408, Michigan State University.

Dahlhaus, R. (1989) "Efficient parameter estimation for self-similar processes". *Annals of Statistics*, 17, 1749-1766.

Davidson, J. (1994) *Stochastic Limit Theory*. Oxford University Press.

Delgado M. and J. Hidalgo (1999) "Bootstrap Goodness of fit tests for FARIMA models". Working Paper 99-38, Universidad Carlos III de Madrid.

Dolado J.J., J. Gonzalo and L. Mayoral (2003) "Testing I(1) against I(d) alternatives in the presence of deterministic components". Mimeo.

Fox, R. and M.S. Taqqu (1986) "Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series". *The Annals of Statistics,* 14, 517-532.

Galbraith, J. W. and V. Zinde-Walsh (1997) "Time domain methods for the estimation of fractionally-integrated time series models". Mimeo.

Granger C.W.J. and K. Joyeux (1980) "An introduction to long-memory time series and fractional differencing". *Journal of Time Series Analysis*, 1, 15-29.

Hall, P., H. L. Koul and B.A. Turlauch, (1997), "Note on convergence rates of semiparametric estimators of dependence index". *The Annals of Statistics,* Vol. 25, No. 4., 1725-1739.

Henry, M. and P. Zaffaroni (2002) "The long range dependence paradigm for Macroeconomics and Finance", in *Long range dependence: Theory and applications*, P. Doukhan, G. Oppenheim and M. Taqqu (ed). Birkhäuser, Boston.

Hong, Y. (1996) "Consistent testing for serial correlation of unknown form". *Econometrica* 64, 837-64.

Hosking, J.R.M. (1996) "Asymptotic distribution of the sample mean, autocovariances and autocorrelations of long memory time series". *Journal of Econometrics*, 73, 261-284.

Hosking, J.R.M. (1981) "Fractional differencing". *Biometrika,* 68, 165-176.

Li, W.K and A.I. McLeod (1986) "Fractional time series modelling". *Biometrika,* 73, 217-221.

Ling, S. and W. K. Li (1997) "On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity". *Journal of the American Statistical Association*, Vol. 92, No. 439, 1184-94.

Ljung, G.M. and G.E.P. Box (1978) "Lack of fit in time series models". *Biometrika* 65, 297-303.

Marinucci, D. and P. M. Robinson (1999) "Alternative forms of Brownian motion". *Journal of Statistical Planning and Inference*, 80, 111-122.

Michelacci, C. and P. Zaffaroni (2000) "Fractional Beta Convergence". *Journal of*

Monetary Economics, 45, 129-153.

Milhoj, A. (1981) "A test of fit in time series models". *Biometrika* 68, 177-87.

Newey W.K. and D. McFaden (1994) "Large sample estimation and hypotheses testing". In *Handbook of Econometrics IV,* North Holland.

Odaki, M. (1993) "On the invertibility of fractionally differenced ARIMA processes". *Biometrika,* 80, 703-709.

Phillips, P.C.B. (1999) "Discrete Fourier transforms of fractional processes". Mimeo.

Phillips and Moon, (1999), "Linear regression limit theory for non-stationary panel data", *Econometrica*, Vol. 67, N°5. p. 1057-1111.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications* , 2d ed. New York: Wiley.

Robinson, P.M. (1995) "Gaussian semiparametric estimation of long range dependence". *The Annals of Statistics* 23, No. 5, 1630-1661.

Robinson, P.M. (1994a) "Time series with strong dependence". In C. Sims (ed.), *Advances in Econometrics*, Sixth World Congress. Cambridge: Cambridge University Press.

Robinson, P.M. (1994b) "Efficient tests of nonstationary hypotheses". *Journal of the American Statistical Association* 89, 1420-1437.

Rudin, W. (1980) *Principles of Mathematical Analysis*, McGraw-Hill.

Sowell, F.B. (1992) "Maximum likelihood estimation of stationary univariate fractionally-integrated time-series models". *Journal of Econometrics* 53, 165-188.

Sowell, F.B. (1990) "The fractional unit root distribution". *Econometrica* 58, No 2, 495-505.

Tanaka, K. (1999) "The nonstationary fractional unit root". *Econometric Theory*, No. 4, Vol. 15, 549-577.

Tieslau, M., P. Schmidt and R. Baillie (1996) "A minimum distance estimator for long-memory processes". *Journal of Econometrics* 71, 249-264.

Velasco C. and P. M. Robinson (2000) "Whittle pseudo-Maximum likelihood estimation for nonstationary time series", *Journal of the American Statistical Association* 95**,** 1229-1243.

WRIGHT, J.H. (1999) "A new estimator of the fractionally integrated stochastic volatility model". *Economics Letters* 63, 295-303.

Appendix 1

Proof of Theorem 1.

The residuals $e_t(\lambda)$ in (7) evaluated at $\lambda = \lambda_0$ can be written as

$$e_t(\lambda_0) = \sum_{j=0}^{t-m_0-1} \alpha_j(\lambda_0^*)(\Delta^{m_0} y_{t-j} - \mu_0) + \left(\overline{\Delta^{m_0} y} - \mu_0\right) \sum_{j=0}^{t-m_0-1} \alpha_j(\lambda_0^*), \qquad t = m_0+1, ..., T.$$

This allows us to write

$$e_t(\lambda_0) = \varepsilon_t + \eta_t,$$

where η_t is given by

$$\eta_t = -\sum_{j=t-m_0}^{\infty} \alpha_j(\lambda_0^*)(\Delta^{m_0} y_{t-j} - \mu_0) + \left(\overline{\Delta^{m_0} y} - \mu_0\right) \sum_{j=0}^{t-m_0-1} \alpha_j(\lambda_0^*),$$

and $\overline{\Delta^{m_0} y}$ denotes the sample mean of $\Delta^{m_0} y_t$. To simplify the notation, assume without loss of generality that $m_0 = 0$ (when this is not the case, define the process $x_t = \Delta^{m_0} y_t$ and the following arguments will remain valid just by substituting y_t by x_t). Then, $d_0 = \varphi_0$.

Let $\sqrt{T}\Upsilon_T$ be the vector that contains the differences in expression (12), that is,

$$\sqrt{T}\Upsilon_T = \sqrt{T}\left(\hat{\rho}_{ke(\lambda_0)} - \hat{\rho}_{k\varepsilon}\right),$$

and consider the $i-th$ element of the vector $\sqrt{T}\Upsilon_T$. The denominator is given by the sample variance of $e_t(\lambda_0)$ that converges to the innovation variance as long as $d_0 > -1$ (see Odaki, 1993). With respect to the numerator, it is given by

$$T^{-1/2}\left(\sum_{t=1}^{T-i} \varepsilon_{t+i}\eta_t + \sum_{t=1}^{T-i} \varepsilon_t \eta_{t+i} + \sum_{t=1}^{T-i} \eta_t \eta_{t+i}\right). \tag{22}$$

Let us first consider the case where μ_0 is known and equal to zero. In this case η_t collapses to $\eta_t = -\sum_{i=t}^{\infty} \alpha_i(\lambda_0^*) y_{t-i}$. By repeated substitution, Odaki (1993) shows that this expression can be rewritten as

$$\eta_t = \sum_{j=0}^{\infty} \psi_{j,t-1}\varepsilon_{-j}, \tag{23}$$

(for the precise form of the sequence of coefficients $\{\psi_{j,t-1}\}_{j=0}^{\infty}$, see Odaki, 1993, p. 704). He also shows that the orders of magnitude of the sum of squares of these coefficients are,

$$\left(\sum_{j=0}^{\infty}\psi_{j,t-1}^{2}\right)=\begin{cases} O\left(t^{-1}\right)\ if\ \varphi_0\in(-0.5,0.5), \\ O\left(\left(\log t\right)t^{-1}\right)\ if\ \varphi_0=-0.5, \\ O\left(t^{-2(1+\varphi_0)}\right)\ if\ \varphi_0<-0.5. \end{cases} \tag{24}$$

Now we check that the three terms in (22) converge in mean square to zero. Noticing that only terms of ε_t with $t\leq 0$ enter the definition of η_t, it follows that $T^{-1/2}\sum_{t=1}^{T-i}E\left(\varepsilon_{t+i}\eta_t\right)=T^{-1/2}\sum_{t=1}^{T-i}E\left(\varepsilon_{t+i}\right)E\left(\eta_t\right)=0$, by independence of the processes η_t and $\{\varepsilon_t\}_{t=1}^{\infty}$. Taking into account (24),[8] it follows that,

$$T^{-1}E\left(\sum_{t=1}^{T-i}\varepsilon_{t+i}^{2}\eta_t^{2}\right)=\sigma^{4}T^{-1}\sum_{t=1}^{T-i}\left(\sum_{j=0}^{\infty}\psi_{j,t-1}^{2}\right)\xrightarrow{p}0, \tag{25}$$

since it follows from (24) that,

$$\sum_{t=1}^{T-i}\left(\sum_{j=0}^{\infty}\psi_{j,t-1}^{2}\right)=\begin{cases} O\left(\log(T-i)\right)\ for\ \varphi_0\in(-0.5,0.5) \\ O\left(\left(\log(T-i)\right)^{2}\right)\ for\ \varphi_0=-0.5 \\ O\left((T-i)^{-2(1+\varphi_0)+1}\right)\ for\ \varphi_0<-0.5. \end{cases} \tag{26}$$

A similar argument holds for the second term in (22), since in this case $\eta_{t+i}=\sum_{j=0}^{\infty}\psi_{j,t+i-1}\varepsilon_{-j}$ and again no contemporaneous terms of ε are found in the product $\varepsilon_t\eta_{t+i}$. With respect to the third term in (22),

$$T^{-1/2}\sum_{t=1}^{T-i}E\left|\eta_t\eta_{t+i}\right|\leq T^{-1/2}\sum_{t=1}^{T-i}E\left(\eta_t^{2}\right)=T^{-1/2}\sigma^{2}\sum_{t=1}^{T-i}\left(\sum_{j=0}^{\infty}\psi_{j,t-1}^{2}\right)\xrightarrow{p}0. \tag{27}$$

To check that the variance also converges to zero, notice that

$$T^{-1}E\left(\sum_{t=1}^{T-i}\eta_t\eta_{t+i}\right)^{2}=2T^{-1}\sum_{r=1}^{T-i}\sum_{s\geq r}^{T-i}E\left(\eta_r\eta_{r+i}\eta_s\eta_{s+i}\right), \tag{28}$$

It can be checked that the coefficients $\{\psi_{j,t}\}$ are strictly smaller than one in absolute value and strictly decreasing in both subindexes (j,t). Using these results and Cauchy's

[8] Since $\sum_{j=0}^{\infty}E\left(\psi_{j,t-1}^{2}\varepsilon_{-j}^{2}\right)<\infty$, it is possible to interchange the expectation and the summation in order to obtain (25). See Rao, (1973), p.111.

inequality, it is obtained that

$$E\left(\eta_r\eta_{r+i}\eta_s\eta_{s+i}\right) \leq \mu_4\left(\sum_{j=0}^{\infty}\psi_{j,r-1}^2\right)^{1/2}\left(\sum_{j=0}^{\infty}\psi_{j,s-1}^2\right)^{1/2} + 3\sigma^4\sum_{j=0}^{\infty}\psi_{j,r-1}^2\sum_{j=0}^{\infty}\psi_{j,s-1}^2. \quad (29)$$

Taking into account the expression above, it is easy to check, by applying the orders of magnitude in (26) , that (28) converges in probability to zero for $\varphi_0 > -0.75$. Since the i-th element of $\sqrt{T}\Upsilon_T$ tends to zero for all $i = 1, ..., k$, then $\sqrt{T}\Upsilon_T \xrightarrow{p} 0$, implying the desired result.

The case where μ_0 is unknown can be proved similarly using standard arguments since $\overline{\Delta^{m_0}y}$ is a consistent estimator of μ_0 (see Robinson, 1994b).■

Proof of Theorem 2

As in the proof of Theorem 1, we assume without loss of generality, that $m_0 = 0$ (otherwise, define the process $x_t = \Delta^{m_0}y_t$ and the following arguments will remain valid just by substituting y_t by x_t). In order to complete the proof, we proceed in several steps. Let $\tilde{\rho}_{k^\varepsilon}$ be the $k \times 1$ vector

$$\tilde{\rho}_{k^\varepsilon} = \left(\begin{array}{ccc} \tilde{\rho}_\varepsilon(1), & ... & ,\tilde{\rho}_\varepsilon(k) \end{array}\right)', \quad (30)$$

whose $j-th$ element is given by $\tilde{\rho}_\varepsilon(j) = \sigma^{-2}\sum_{t=1}^{T-j}\varepsilon_t\varepsilon_{t+j}/T$, that is, is analogous to the vector of innovation autocorrelations, $\hat{\rho}_{k^\varepsilon}$, defined in Theorem 1, but the sample innovation variance has been replaced by the true variance, σ^2. We first show that the vector $\sqrt{T}A'_{k(T)}\tilde{\rho}_{k(T)^\varepsilon}$ converges in distribution to a $N_h(0,\Omega)$ distribution. Then, it will be easy to extend the result to the vector $\sqrt{T}\hat{A}'_{k(T)}\hat{\rho}_{k(T)e(\lambda_0)}$.

From Rao (1973), $T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)^\varepsilon}$ jointly converge to the multivariate normal in (15) iff for any $h-$dimensional vector of real numbers $\omega = (\omega_1, ..., \omega_h)'$, $\sqrt{T}\omega'A'_{k(T)}\tilde{\rho}_{k(T)^\varepsilon}$ converges in distribution to $N(0, \omega'\Omega\omega)$. Let $T^{1/2}\sum_{i=1}^{k(T)}a_{ij}\tilde{\rho}_\varepsilon(i)$ be the $j-th$ element of the vector $T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)^\varepsilon}$. We next consider the limit of

$$T^{1/2}\sum_{i=1}^{k(T)}\left(\omega_1a_{i1} + \omega_2a_{i2} + ... + \omega_ha_{ih}\right)\tilde{\rho}_\varepsilon(i), \quad (31)$$

for any ω. To simplify the notation, let us call $\mathfrak{A}_i = \omega_1a_{i1} + \omega_2a_{i2} + ... + \omega_ha_{ih}$. Robinson (1994, p. 1434) derives the distribution of $T^{1/2}\sum_{i=1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)$ for the case where $k(T) =$

$T-1$. For the sake of completeness, we sketch now his proof. For a fixed m, it is possible to write (31) as,

$$T^{1/2}\sum_{i=1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)=T^{1/2}\sum_{i=1}^{m}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)+T^{1/2}\sum_{i=m+1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i). \tag{32}$$

Application of a martingale difference central limit theorem (e.g. Brown, 1971) implies that, for fixed m, $T^{1/2}\sum_{i=1}^{m}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)\xrightarrow{w}N(0,\sum_{i=1}^{m}\mathfrak{A}_i^2)$. By Bernstein lemma (see Bernstein, 1927), if

$$\lim_{T,k(T)\to\infty}E\left(\left(T^{1/2}\sum_{i=m+1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)\right)^2\right)\to 0, \tag{33}$$

then $\lim_{T\to\infty}T^{1/2}\sum_{i=1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i)\xrightarrow{w}N(0,\sum_{i=1}^{\infty}\mathfrak{A}_i^2)$. To show that (33) holds, notice that

$$TE(\sum_{i=m+1}^{k(T)}\mathfrak{A}_i\tilde{\rho}_\varepsilon(i))^2=O(\sum_{i=m+1}^{\infty}\mathfrak{A}_i^2),$$

since $E\left(\tilde{\rho}_\varepsilon(i)^2\right)=(T-i)/T^2$ and $E\left(\tilde{\rho}_\varepsilon(j)\,\tilde{\rho}_\varepsilon(i)\right)=0$ for $i\neq j$, by independence of the $\varepsilon's$. Finally,

$$\begin{aligned}\sum_{i=1}^{\infty}\mathfrak{A}_i^2 &= \sum_{i=1}^{\infty}(\omega_1a_{i1}+\omega_2a_{i2}+...+\omega_ha_{ih})^2\\ &= \sum_{j=1}^{h}\omega_h^2\left(\sum_{i=1}^{\infty}a_{ij}^2\right)+\sum_{s=1}^{h}\sum_{l=1}^{h}\omega_s\omega_l\left(\sum_{i=1}^{\infty}a_{is}a_{il}\right)=\omega'\Omega\omega<\infty,\end{aligned}$$

since Ω is a matrix of (finite) real numbers. This means that $\sum_{i=m+1}^{\infty}\mathfrak{A}_i^2$ can be made arbitrarily small by choosing m large enough, implying that (33) holds and, in turn, that

$$lim_{T,k(T)\to\infty}\sqrt{T}A'_{k(T)}\tilde{\rho}_{k(T)\varepsilon}\xrightarrow{w}N_h(0,\Omega). \tag{34}$$

The next step is to show that the vector $\tilde{\rho}_{k\varepsilon}$ can be replaced in (34) by $\tilde{\rho}_{ke(\lambda_0)}=\left(\tilde{\rho}_{e(\lambda_0)}(1),\ ...,\ \tilde{\rho}_{e(\lambda_0)}(k)\right)'$, where

$$\tilde{\rho}_{e(\lambda_0)}(i)=\sigma^{-2}\sum_{t=1}^{T-i}e_t(\lambda_0)\,e_{t+i}(\lambda_0)/T. \tag{35}$$

Consider,

$$T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)e(\lambda_0)}=T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)\varepsilon}+T^{1/2}A'_{k(T)}(\tilde{\rho}_{k(T)e(\lambda_0)}-\tilde{\rho}_{k(T)\varepsilon}). \tag{36}$$

By the Slutsky's theorem, if the second term in the RHS of (36) converges in probability to zero, then the distribution of $T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)e(\lambda_0)}$ and $T^{1/2}A'_{k(T)}\tilde{\rho}_{k(T)^{\varepsilon}}$ is the same. For the sake of simplicity, assume that $h = 1$ (the multivariate case can be solved as in the lines above). By Definition 2 in Phillips and Moon, (1999) if

$$\lim_{T,k\to\infty} P\left(\left|T^{1/2}\sum_{i=1}^{k} a_{i1}\left(\tilde{\rho}_{e(\lambda_0)}(i) - \tilde{\rho}_{\varepsilon}(i)\right)\right| > \epsilon\right) = 0,\ \forall\epsilon > 0 \tag{37}$$

holds, then the limit of $T^{1/2}A'_{k(T)}(\tilde{\rho}_{k(T)e(\lambda_0)} - \tilde{\rho}_{k(T)^{\varepsilon}})$ when $T,\ k \to \infty$ is equal to zero. Noticing that

$$T^{1/2}\sum_{i=1}^{k} a_{i1}\left(\tilde{\rho}_{e(\lambda_0)}(i) - \tilde{\rho}_{\varepsilon}(i)\right) \leq T^{1/2}\left(\sum_{i=1}^{k} a_{i1}^2\right)^{1/2}\left(\sum_{i=1}^{k}\left(\tilde{\rho}_{e(\lambda_0)}(i) - \tilde{\rho}_{\varepsilon}(i)\right)^2\right)^{1/2},$$

it follows that the probability in (37) can be bounded by,

$$\epsilon^{-2}\left(\sum_{i=1}^{k} a_{i1}^2\right) TE\left(\sum_{i=1}^{k}\left(\tilde{\rho}_{e(\lambda_0)}(i) - \tilde{\rho}_{\varepsilon}(i)\right)^2\right). \tag{38}$$

From (22),

$$T\sum_{i=1}^{k} E(\tilde{\rho}_{e(\lambda_0)}(i) - \tilde{\rho}_{\varepsilon}(i))^2 = T^{-1}\sigma^{-4}\sum_{i=1}^{k}\left\{E\left(\sum_{t=1}^{T-i}\varepsilon_t\eta_{t+i} + \varepsilon_{t+i}\eta_t\right)^2\right. \tag{39}$$

$$\left. +2E\left(\sum_{t=1}^{T-i}(\eta_t\eta_{t+i})\sum_{t=1}^{T-i}(\varepsilon_t\eta_{t+i} + \varepsilon_{t+i}\eta_t)\right) + E\left(\sum_{t=1}^{T-i}\eta_t\eta_{t+i}\right)^2\right\}. \tag{40}$$

Assume now for simplicity that μ_0 is known and thus, η_t is defined as in (23). We now check that the expression above tends to zero. By independence of the $\varepsilon's$ and the results in the proof of Theorem 1, it is easy to obtain that,

$$T^{-1}\sigma^{-4}\sum_{i=1}^{k} E\left(\sum_{t=1}^{T-i}\varepsilon_t\eta_{t+i} + \varepsilon_{t+i}\eta_t\right)^2 \leq 2T^{-1}\sum_{i=1}^{k}\left(\sum_{t=1}^{T-i}\left(\sum_{s=0}^{\infty}\psi_{s,t-1}^2\right)\right) \tag{41}$$

If $d_0 > -0.5$, by (26) it is known that $\sum_{t=1}^{T-i}\left(\sum_{j=0}^{\infty}\psi_{j,t-1}^2\right) = O(\log(T-i))$. Then, it follows that $T^{-1}\sum_{i=1}^{k}\log(T-i)$ is $O\left(\frac{(T-k)(1-\log(T-k))-T(1-\log(T))}{T}\right)$ and applying L'Hôpital's rule, it can be checked that this expressions tends to zero as long as $k/T \to 0$. If -0.75<

$d_0 \leq -0.5$, application of a similar argument, using the corresponding orders of magnitude in (26), guarantee that (41) tends to zero for all $d_0 > -0.75$. The first term in (40) is equal to zero by independence of the $\varepsilon' s$, and therefore its limit is trivially equal to zero. Taking into account (26), (28) and (29) and following the same strategy as in the lines above, it is straight forward to check that the second term in (40) also tends to zero for $d_0 > -0.75$. Finally, since $\lim_{k,T\to\infty}\left(\sum_{i=1}^{k} a_{i1}^2\right) = \sum_{i=1}^{\infty} a_{i1}^2 < \infty$ by assumption, it follows that (38) tends to zero. Consequently $T^{1/2} A'_{k(T)} \tilde{\rho}_{k(T)e(\lambda_0)} \xrightarrow{d} N(0,\Omega)$.

Furthermore, application of the Slutsky's Theorem allows to substitute σ^2 by its sample counterpart, $\sum e_t^2(\lambda_0)/T$, implying that $T^{1/2} A'_k \hat{\rho}_{e(\lambda_0)}(j) \xrightarrow{d} N(0,\Omega)$.

The final step is to substitute $A_{k(T)}$ by $\hat{A}_{k(T)}$. Since,

$$T^{1/2}\hat{A}'_{k(T)}\hat{\rho}_{k(T)e(\lambda_0)} = T^{1/2}A'_{k(T)}\hat{\rho}_{k(T)e(\lambda_0)} + T^{1/2}(\hat{A}'_{k(T)} - A'_{k(T)})\hat{\rho}_{k(T)e(\lambda_0)}, \tag{42}$$

then, proving that the second term in the RHS of (42) converges to zero when $T, k \to \infty$, implies that $T^{1/2}\hat{A}'_{k(T)}\hat{\rho}_{k(T)e(\lambda_0)}$ and $T^{1/2}A'_{k(T)}\hat{\rho}_{k(T)e(\lambda_0)}$ share the same asymptotic distribution. This is equivalent to show that the limit of $T^{1/2}\sum_{i=1}^{k(T)}(\hat{a}_{ij} - a_{ij})\hat{\rho}_{e(\lambda_0)}(i)$ when both T and k go to infinity is equal to zero for $j = 1, ..., h$. Notice that,

$$\begin{aligned} T^{1/2}\sum_{i=1}^{k(T)}(\hat{a}_{ij} - a_{ij})\hat{\rho}_{e(\lambda_0)}(i) &= T^{1/2}\sum_{i=1}^{k(T)}\left(\frac{\hat{a}_{ij} - a_{ij}}{a_{ij}}\right)a_{ij}\hat{\rho}_{e(\lambda_0)}(i), \\ &\leq \left(\sup_{1\leq i\leq k}\left(\frac{\hat{a}_{ij}}{a_{ij}}\right) - 1\right)T^{1/2}\sum_{i=1}^{k(T)} a_{ij}\hat{\rho}_{e(\lambda_0)}(i) \xrightarrow{p} 0 \end{aligned} \tag{43}$$

for all $j = 1, ..., h$, since the first factor of (43) goes to zero by assumption while the second has been shown above to be $O_p(1)$, implying that $\sqrt{T}\hat{A}'_k\hat{\rho}_{ke(\lambda_0)} \xrightarrow{d} N(0,\Omega)$, as desired.■

In order to prove the consistency of the estimator stated in Theorem 3, we consider separately the cases where the inferior limit of the parameter space of d, ∇_1, is such that $\nabla_1 \leq d_0 - 1/2$ from those where $\nabla_1 > d_0 - 1/2$. The reason for making this distinction is the non-uniform behavior of $FI(d)$ processes which determines the properties of the criterion function. More specifically, whenever d is $\geq 1/2$, correlations are not well-defined since the process is not stationary. To define the residuals, $e_t(d)$, the process y_t is filtered by Δ^d, so

that $e_t(d)$ is a $FI(d_0 - d)$ process. If $\nabla_1 > d_0 - 1/2$, then $d_0 - d < 1/2$ for all $d \in [\nabla_1, \nabla_2]$, and the autocorrelations are well-defined. But if $\nabla_1 \leq d_0 - 1/2$, the difference $(d_0 - d)$ could be bigger or smaller than $1/2$ so that $e_t(d)$ could be or not be (asymptotically) stationary, depending on the value of d. A solution to this problem appeared in Robinson (1995) and Velasco and Robinson (2000). They proposed a two-step proof that will be also followed in this article. The proof is based on dividing the whole parametric space into two subsets: the first containing only the values of the parameters for which the filtered process is stationary and the second, the remaining ones. The first step of the proof shows that the estimator computed only considering the first subset is consistent, while the second proves that values in the second subset cannot be (asymptotically) optimal, which implies that optimal values are always in the first subset and, consequently, that the estimator computed in the whole parametric space is consistent.

Before stating the proof, we need the following definition.

Definition 1 *Denote* $\lambda^{(1)} = d$ *and* $\Lambda_1 = \{d : \nabla_1 \leq d \leq \nabla_2\} \times \Lambda^{(-1)}$, *if* $\nabla_1 > d_0 - 1/2$ *or otherwise* $\Lambda_1 = \{d : d_0 - 1/2 + \eta \leq d \leq \nabla_2\} \times \Lambda^{(-1)}$ *for some* $\eta \in (0, 1/2)$, *if* $\nabla_1 \leq d_0 - 1/2$, *where* $\Lambda^{(-1)}$ *is the parameter space of the remaining ARMA parameters.*

The following auxiliary result is needed.

Lemma 1 *Let* $V_{ke}(\lambda)$ *be the criterion function in* (13), *where* $e_t(\lambda)$ *is defined as in* (5) *or* (7), *according to the case where the DGP has a known or unknown mean respectively. Let* $\tilde{\lambda} \in \Lambda_1$ *and define* $V_{k\varepsilon}\left(\tilde{\lambda}\right) = \sum_{i=1}^{k}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2$, *where* $\rho_{\varepsilon(\tilde{\lambda})}(.)$ *are the (population) autocorrelations associated to the non truncated residuals* $\varepsilon\left(\tilde{\lambda}\right) = \sum_{j=0}^{\infty}\alpha_j\left(\tilde{\lambda}\right)x(m)_{t-j}$, *and* k *is a fixed number. Then:*

1. $V_{ke}(\lambda)$ *is continuous in* λ, $V_{ke}\left(\tilde{\lambda}\right)$ *converges in probability to* $V_{k\varepsilon}\left(\tilde{\lambda}\right)$ *and the convergence is uniform.*

2. $V_{k\varepsilon}\left(\tilde{\lambda}\right)$ *is a continuous function and has a unique minimum at* λ_0, *such that* $V_{k\varepsilon}(\lambda_0) = 0$.

Proof of Lemma 1.

1. The continuity of $V_{ke}(\lambda)$ is trivial, since it is a continuous composition of continuous functions. The asymptotic negligibility of the truncation can be proved along the lines of Theorem 1. Since the sample correlations associated to stationary processes are consistent (Hosking, 1996), it follows that $V_{ke}\left(\tilde{\lambda}\right) \xrightarrow{p} V_{k\varepsilon}\left(\tilde{\lambda}\right)$ for any $\tilde{\lambda} \in \Lambda_1$. The uniform convergence follows from the pointwise convergence and an equicontinuity argument using the compactness of Λ_1 and the differentiability of $\rho(.)$ with respect to λ (cf. Davidson, (1994), p. 340, and Velasco and Robinson, 2000).

2. It is straightforward to check that it has a unique minimum at λ_0, since $\varepsilon_t(\lambda)|_{\lambda=\lambda_0} = \varepsilon_t$, which is an *i.i.d.* process and therefore all its correlations are zero, (which implies that $V_{k\varepsilon}(\lambda_0) = 0$), but presents non-null autocorrelations for any other value of $\tilde{\lambda} \neq \lambda_0$ (and therefore $V_{\varepsilon}\left(\tilde{\lambda}\right) > 0$). The continuity of $V_{k\varepsilon}\left(\tilde{\lambda}\right)$ follows from the assumptions above (see Amemiya, 1985, Theorem 4.1.1)■

Proof of Theorem 3

We proceed with the two-step proof of consistency proposed in Robinson (1995) and Velasco and Robinson (2000).

First step. Define $\hat{\lambda}_1 = \arg\min_{\lambda\in\Lambda_1} V_{ke}(\lambda)$. It follows from standard results that if we can write $V_{ke}(\lambda) - V_{ke}(\lambda_0) = S(\lambda) - U(\lambda)$, where $S(\lambda)$ is nonstochastic, constant over t and for all $\epsilon > 0$ there exists $\delta > 0$ such that $\inf_{\|\lambda-\lambda_0\|\geq\epsilon} S(\lambda) \geq \delta$, and $\sup_{\lambda\in\Lambda_1} |U(\lambda)| \xrightarrow{p} 0$, then $\hat{\lambda}_1 \xrightarrow{p} \lambda_0$. Hence, let us denote $S(\lambda) = V_{k\varepsilon}(\lambda)$ and, since $V_{k\varepsilon}(\lambda)$ is continuous and has a unique minimum at $\lambda = \lambda_0$, the conditions on $S(\lambda)$ hold; $U(\lambda)$ is given in turn by $U(\lambda) = V_{k\varepsilon}(\lambda) - V_{ke}(\lambda) + V_{ke}(\lambda_0)$. Notice that:

$$\sup_{\lambda\in\Lambda_1} |U(\lambda)| \leq \sup_{\lambda\in\Lambda_1} |V_{ke}(\lambda) - V_{k\varepsilon}(\lambda)| + |V_{ke}(\lambda_0)|, \tag{44}$$

and both terms in the right hand side of (44) tend to zero, the first due to uniform convergence and the second due to pointwise convergence (Lemma 1).

Second step. Recall that $\Lambda_1 = \{d : \nabla \leq d \leq \nabla_2\} \times \Lambda^{(-1)}$, where $\nabla = \nabla_1$, if $d_0 < \nabla_1 + 1/2$ and $d_0 - 1/2 < \nabla < d_0$ otherwise. If $d_0 < \nabla_1 + 1/2$, then $\Lambda_1 = \Lambda$ and the theorem is proved. When $d_0 > \nabla_1 + 1/2$, define $\Lambda_2 = \{d : \nabla_1 \leq d < d_0 - 1/2 + \eta\} \times \Lambda^{(-1)}$ and let $\hat{\lambda}_k = \arg\min_{\lambda \in \Lambda} V_{ke}(\lambda)$ be the estimator computed in the whole parametric space, $\Lambda = \Lambda_1 \cup \Lambda_2$. We want to show that $\hat{\lambda}_k - \hat{\lambda}_1 \xrightarrow{p} 0$ or, equivalently, that for any $\delta > 0$, $P\left(\left\|\hat{\lambda}_k - \hat{\lambda}_1\right\| \geq \delta\right) \to 0$. Notice that,

$$\begin{aligned} P\left(\left\|\hat{\lambda}_k - \hat{\lambda}_1\right\| \geq \delta\right) &\leq P\left(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq \min_{\lambda \in \Lambda_1} V_{ke}(\lambda)\right) \\ &= P\left(\left(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq V_{ke}\left(\hat{\lambda}_1\right)\right) \cap \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0\right) \qquad (45) \\ &\quad + P\left(\left(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq V_{ke}\left(\hat{\lambda}_1\right)\right) \cap \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) > 0\right) \\ &\leq P\left(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0\right) + P\left(V_{ke}\left(\hat{\lambda}_1\right) > 0\right). \qquad (46) \end{aligned}$$

Since $\hat{\lambda}_1$ is consistent (step 1) and uniform convergence of V_{ke} holds, the second probability in (46) tends to zero. To check that the first one also tends to zero, recall that the function $V_{ke}(\lambda)$, $\lambda \in \bar{\Lambda}_2$, where $\bar{\Lambda}_2$ is the closure of Λ_2, contains the squared sample correlations of a $FI\left(d_0 - \lambda^{(1)}\right)$ process, where $\lambda^{(1)} \in [\nabla_1, d_0 - 1/2 + \eta]$, therefore it is always a non-negative quantity. Whenever $\lambda^{(1)} \in (d_0 - 1/2, d_0 - 1/2 + \eta]$, then $1/2 - \eta < d_0 - \lambda^{(1)} < 1/2$, and the corresponding filtered process is long-memory stationary. Thus, the squared sample autocorrelations converge to the squared population autocorrelations which, clearly, are bounded away from zero. If $\lambda^{(1)} \in [\nabla_1, d_0 - 1/2]$, then $d_0 - \lambda^{(1)} \geq 1/2$, so that $V_{ke}(\lambda)$ contains the sample autocorrelations of a non-stationary process. Since $V_{ke}(\lambda)$ should contain, at least, the first autocorrelation, it is clear that $V_{ke}(\lambda) = \sum_{i=1}^{k} \hat{\rho}^2_{e(\lambda)}(i) \geq \hat{\rho}^2_{e(\lambda)}(1) \xrightarrow{p} 1$ (see Sowell, (1990), Theorem 3). The continuity of $V_{ke}(\lambda)$ implies that the infimum is contained in $\bar{\Lambda}_2$. Therefore it holds that $P\left(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0\right) \to 0$.■

In order to derive the asymptotic distribution of $\hat{\lambda}_k$, the following auxiliary result will be useful.

Lemma 2 *Let $\hat{J}_k(\lambda) = \partial \hat{\rho}_{ke(\lambda)} / \partial \lambda'$ be the Jacobian matrix of $\hat{\rho}_{ke(\lambda)}$, that is, the $k \times (p+q+1)$ matrix of partial derivatives of $\hat{\rho}_{ke(\lambda)}$ with respect to λ. Under the hypotheses of Sections*

2 and 3, it holds that,

1. *For each finite k,*

$$\hat{J}_k(\lambda_0) \xrightarrow{p} J_k(\lambda_0),$$

where,

$$J_k(\lambda_0) = \begin{pmatrix} -1 & 1 & 0 & ... & 1 & ... & 0 \\ -1/2 & \omega_1 & 1 & ... & \psi_1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ -1/k & \omega_{k-1} & \omega_{k-2} & ... & \psi_{k-1} & ... & \psi_{k-q} \end{pmatrix}. \tag{47}$$

and the coefficients ω_i and ψ_i are defined by the equations $\sum \omega_i L^i = \frac{1}{\Phi(L)}$ and $\sum \psi_i L^i = \frac{1}{\Theta(L)}$.

2. *The first derivative of $\hat{J}_k(\lambda)$ exists and is bounded in probability in an open convex set containing λ_0.*

Proof of Lemma 2

1. This result can be found in Theorem 2 in Li and McLeod (1986).

2. Let $\tilde{\Lambda} \subset \Lambda_1$ be an open convex set containing λ_0. Notice that $\hat{J}_k(\lambda)$ is differentiable in this set since it is a function of the derivatives of the residuals $e_t(\lambda)$ which are $C(\infty)$ with respect to λ in the relevant set. The first derivative of $\hat{J}_k(\lambda)$ is the $k(p+q+1) \times (p+q+1)$ matrix defined as,

$$\partial vec(\partial \hat{\rho}_{ke(\lambda)}/\partial \lambda')/\partial \lambda'.$$

Consider the $(1,1)$ element of this matrix, given by $\partial^2 \hat{\rho}_{e(\lambda)}(1)/\partial d^2$. To simplify the notation, we use the symbols $'$ and $''$ to denote first and second derivative, respectively, with respect to d. It follows that,

$$\hat{\rho}''_{e(\lambda)}(1) = \frac{\sigma^2_{e(\lambda)}}{\sum e_t^2(\lambda)/T} \left\{ \tilde{\rho}''_{e(\lambda)}(1) - \tilde{\rho}'_{e(\lambda)}(1) \frac{4\sum e_t e'_t(\lambda)}{\sum e_t^2(\lambda)} \right. \tag{48}$$

$$\left. -\tilde{\rho}_{e(\lambda)}(1) \left(\frac{2\left(\sum e_t e''_t(\lambda) + \sum (e'_t(\lambda))^2\right)}{\sum e_t^2(\lambda)} - \frac{8\left(\sum e_t e'_t(\lambda)\right)}{\left(\sum e_t^2(\lambda)\right)^2} \right) \right\}, \tag{49}$$

where $\tilde{\rho}_{e(\lambda)}(1)$ is defined as in (35), $\sigma^2_{e(\lambda)} = \lim_{T\to\infty} T^{-1}\sum e_t^2(\lambda)$,

$$\tilde{\rho}'_{e(\lambda)}(1) = \frac{\sum e'_t(\lambda)\, e_{t+1}(\lambda) + \sum e_t(\lambda)\, e'_{t+1}(\lambda)}{T},$$

and

$$\tilde{\rho}''_{e(\lambda)}(1) = \frac{\sum e''_{t+1}(\lambda)\, e_t(\lambda) + \sum e''_t(\lambda)\, e_{t+1}(\lambda) + 2\sum e'_t(\lambda)\, e'_{t+1}(\lambda)}{T}. \tag{50}$$

Noticing that $e_t = (1-L)^d \Theta^{-1}\Phi(L)\, y_t 1_{(t>0)}$, it follows that $\partial e_t/\partial d = \log(1-L)e_t$, and $\partial^2 e_t/\partial d^2 = \log^2(1-L)e_t$. Taking into account the expansions $\log(1-L) = -\left(L + \frac{L^2}{2} + \frac{L^3}{3} + ...\right)$ and $\log^2(1-L) = \left(\varkappa_2 L^2 + \varkappa_3 L^3 + ...\right)$ with $\varkappa_i L^i = \frac{2}{i}\left(\sum_{j=1}^{i-1} j^{-1}\right) L^i$ for $i = 0, 1, ...$ and that $e_t = 0$ for all $t < 0$, it follows that,

$$e'_t(\lambda) = \sum_{h=1}^{t-1} \frac{e_{t-h}}{h} \text{ and } e''_t(\lambda) = \sum_{h=2}^{t-1} \varkappa_h e_{t-h}.$$

It is easy, although quite lengthy, to check that each of terms in the RHS of (48) and (49) are $O_p(1)$. Consider for instance $\tilde{\rho}''_{e(\lambda)}(1)$. The first term in (50) can be rewritten as,

$$\frac{\sum\left(\log^2(1-L)\, e_{t+1}(d)\right) e_t(d)}{T} = \frac{\sum\left[\varkappa_2 e_{t-1}(d) + \varkappa_3 e_{t-2}(d) + \varkappa_4 e_{t-3}(d) + ... + e_1(d)\right] e_t(d)}{T}$$

$$= \varkappa_2\hat{\gamma}_e(1) + \varkappa_3\hat{\gamma}_e(2) + \varkappa_4\hat{\gamma}_e(3) + ... + \varkappa_{T-1}\hat{\gamma}_e(T-2)$$

where $\hat{\gamma}_e(.)$ is the sample covariance function of the process $e_t(d)$. We now check that this sum is finite when T tends to ∞. Since $d \in \Lambda_1$, the autocovariance function of $e_t(d)$, $\gamma_e(.)$, decays at a rate $j^{2(d_0-d)-1}$ for large j (see Baillie, 1996). Then,

$$\varkappa_{j+1}\gamma(j) \approx C\frac{2}{j+1}\left(\sum_{i=1}^{j} i^{-1}\right)(j^{2(d_0-d)-1}) \approx C' j^{2(d_0-d)-2}\log j$$

for large j, where C and C' are some constants. Since $(d_0 - d) < 0.5$, then $\sum_{j=1}^{\infty} j^{2(d_0-d)-2}\log j < \infty$. This implies that $\lim_{T\to\infty}\sum_{j=1}^{T-2}\varkappa_{j+1}\gamma(j)$ is bounded. On the other hand, since $\hat{\gamma}(j)$ is a consistent estimator of the covariance function, then

$$\lim_{T\to\infty}\sum_{j=1}^{T-2}\varkappa_{j+1}\hat{\gamma}(j) = O_p(1).$$

The proof for the remaining components in (48) and (49) is analogous and therefore is omitted. Similarly, proofs for the remaining elements of the matrix $\partial vec(\partial \hat{\rho}_{_{k}e(\lambda)}/\partial \lambda')/\partial \lambda'$ can be constructed along the same lines and for the sake of brevity, are also omitted.■

Proof of Theorem 4

The mean value theorem applied to the first-order condition gives,

$$0 = \frac{\partial V_{ke}\left(\lambda_0\right)}{\partial \lambda} + \frac{\partial^2 V_{ke}\left(\bar{\lambda}\right)}{\partial \lambda \partial \lambda'}\left(\hat{\lambda}_k - \lambda_0\right),$$

where $\bar{\lambda}$ belongs to the line joining $\hat{\lambda}_k$ and λ_0. Multiplying through by $\sqrt{T}$ and solving for $\sqrt{T}\left(\hat{\lambda}_k - \lambda_0\right)$ yields

$$\sqrt{T}\left(\hat{\lambda}_k - \lambda_0\right) = -\left(\frac{\partial \hat{\rho}'_{_{k}e(\bar{\lambda})}}{\partial \lambda}\frac{\partial \hat{\rho}_{_{k}e(\bar{\lambda})}}{\partial \lambda'}\right)^{-1}\frac{\partial \hat{\rho}'_{_{k}e(\lambda_0)}}{\partial \lambda}\sqrt{T}\hat{\rho}_{_{k}e(\lambda_0)}. \tag{51}$$

$$= -\left(\hat{J}'_k\left(\bar{\lambda}\right)\hat{J}_k\left(\bar{\lambda}\right)\right)^{-1}\hat{J}_k\left(\lambda_0\right)\sqrt{T}\hat{\rho}_{_{k}e(\lambda_0)}. \tag{52}$$

Notice that $\bar{\lambda}$ can be replaced in (52) by λ_0, since $\bar{\lambda} \xrightarrow{p} \lambda_0$ (due to the consistency of $\hat{\lambda}_k$) and the first derivative of $\hat{J}_k\left(\lambda\right)$ exists and is bounded in probability in an open convex set containing λ_0 (Lemma 2), see Amemiya (1985), Theorem 4.1.4.

Also, since $\sqrt{T}\hat{\rho}_{_{k}e(\lambda_0)} = \sqrt{T}\hat{\rho}_{_{k}\varepsilon} + o_p\left(1\right) \xrightarrow{w} N\left(0, I_k\right)$ for fixed k (Theorem 1), it is straightforward to check that $\sqrt{T}\left(\hat{\lambda}_k - \lambda_0\right)$ is also asymptotically normally distributed with mean equal to zero and variance-covariance matrix given by $\Xi_k^{-1} = \left(J'_k\left(\lambda_0\right)J_k\left(\lambda_0\right)\right)^{-1}$.■

Before proving Theorem 5, the sets Λ_1 and Λ_2 need to be redefined. The reason for introducing this modification is that the criterion function contains in this case an increasing number of squared sample correlations (since $k \to \infty$) and the summability of the squared correlations of $FI\left(\delta\right)$ processes depends upon the value of δ. More specifically, whenever δ is $\geq 1/4$, the corresponding correlations are not square-summable. The two-step proof of consistency introduced in Theorem 3 will also be used in this context but the set Λ_1 will be replaced by Λ_1^*. This set is such that for any $d^* \in \Lambda_1^*$ used to compute the residuals, it holds that $d_0 - d^* < 1/4$, implying that the associated squared autocorrelations are summable.

Definition 2 *Denote* $\lambda^{(1)} = d$ *and* $\Lambda_1^* = \{d : \nabla_1 \leq d \leq \nabla_2\} \times \Lambda^{(-1)}$ *if* $\nabla_1 > d_0 - 1/4$ *or*

otherwise $\Lambda_1^* = \{d : d_0 - 1/4 + \eta \leq d \leq \nabla_2\} \times \Lambda^{(-1)}$ *for some* $\eta \in (0, 1/4)$, *if* $\nabla_1 \leq d_0 - 1/4$, *where* $\Lambda^{(-1)}$ *is the parameter space of the remaining ARMA parameters.*

In addition, the following auxiliary results are needed to prove the theorem.

Lemma 3 *Let* $V_{ke}(\lambda)$ *be the criterion function in* (13), *where* $e_t(\lambda)$ *is defined as in* (5) *or* (7) *according to the case where the DGP has a known or unknown mean respectively and let* $\tilde{\lambda} \in \Lambda_1^*$. *Also define* $V_\varepsilon\left(\tilde{\lambda}\right) = \sum_{i=1}^{\infty}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2$, *where* $\rho_{\varepsilon(\tilde{\lambda})}(.)$ *are the (population) autocorrelations associated to the non truncated residuals* $\varepsilon\left(\tilde{\lambda}\right)$. *Let* k *be a function of* T, *such that* $\lim_{T\to\infty} k(T) = \infty$ *and* $\lim_{T\to\infty} k(T)/T = 0$. *Then:*

1. $V_\varepsilon\left(\tilde{\lambda}\right)$ *is a continuous function and has a unique minimum at* λ_0, *such that* $V_\varepsilon(\lambda_0) = 0$.

2. $V_{ke}\left(\tilde{\lambda}\right)$ *is continuous in* $\tilde{\lambda} \in \Lambda_1^*$, *converges in probability to* $V_\varepsilon\left(\tilde{\lambda}\right)$ *and the convergence is uniform.*

Proof of Lemma 3

1. To check the continuity of $V_\varepsilon\left(\tilde{\lambda}\right)$ it suffices to show that the partial sum of squared autocorrelations, $s_n = \sum_{i=1}^{n}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2$, converges uniformly to $V_\varepsilon\left(\tilde{\lambda}\right)$ (see Rudin, 1980, Theorem 7.12). Applying the Cauchy criterion, this would be true iff for all $\xi > 0$, there exists an integer N such that $m \geq N$, $n \geq N$ and $\tilde{\lambda} \in \Lambda_1^*$, for which

$$\left|\sum_{i=1}^{m}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2 - \sum_{i=1}^{n}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2\right| \leq \xi. \tag{53}$$

   Assume, without loss of generality, that $m > n$. Then, the left-hand side of (53) is simply $\sum_{n+1}^{m}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2$. Since $V_\varepsilon\left(\tilde{\lambda}\right) < \infty$ for all $\varepsilon\left(\tilde{\lambda}\right)$ (recall that $V_\varepsilon\left(\tilde{\lambda}\right)$ contains the squared autocorrelations of FI processes with an order of integration given by $\left(d_0 - \tilde{d}\right)$ strictly smaller of 1/4), there exists an integer N such that for $n > N$, $\sum_{n+1}^{\infty}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2 \to 0$, which clearly implies that $\sum_{i=n+1}^{m}\left(\rho_{\varepsilon(\tilde{\lambda})}(i)\right)^2$ also tends to zero for all $\tilde{\lambda} \in \Lambda_1^*$.

It is straightforward to check that it has a unique minimum at λ_0, since $\varepsilon_t(\lambda)|_{\lambda=\lambda_0} = \varepsilon_t$, which is an *i.i.d.* process and therefore all its correlations are zero, (which implies that $V_\varepsilon(\lambda_0) = 0$), but presents non-zero autocorrelations for any other value of $\tilde{\lambda} \neq \lambda_0$ (and therefore $V_\varepsilon\left(\tilde{\lambda}\right) > 0$).

2. Continuity follows along the same lines as in Lemma 1. The convergence in probability is first proven for the autocovariance function and the result can be easily extended to the autocorrelation function. By Lemma 6 in Phillips and Moon (1999), if *a)* $\sum_{i=1}^{k}\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right)^2 \xrightarrow{p} \sum_{i=1}^{k}\left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2$ for fixed k, *b)* $\lim_{k\to\infty}\sum_{i=1}^{k}\left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2 = \sum_{i=1}^{\infty}\left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2$, and *c)*

$$\lim_{k,T\to\infty}\sup P\left(\left|\sum_{i=1}^{k}\left\{\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right)^2 - \left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2\right\}\right| > \epsilon\right) = 0. \tag{54}$$

then, $\sum_{i=1}^{k}\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right)^2$ converges in probability to $\sum_{i=1}^{\infty}\left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2$ when both k and T are allowed to go to infinity. *a)* follows from Lemma 1; *b)* follows from noticing that the limit is well-defined since autocovariances (and also autocorrelations) of $FI(d)$ processes are square-summable if $d < 1/4$, which is verified here. Finally, by Markov's inequality, and adding and substracting $\left(E\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right)\right)^2$, it follows that the probability in (54) is bounded by

$$\epsilon^{-1}\left(\sum_{i=1}^{k} var\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right) + \sum_{i=1}^{k}\left|\left(E\left(\hat{\gamma}_{e(\tilde{\lambda})}(i)\right)\right)^2 - \left(\gamma_{\varepsilon(\tilde{\lambda})}(i)\right)^2\right|\right). \tag{55}$$

From the results in Hosking (1996) it follows that the variance of the sample autocovariances of $FI(d)$ processes with $d < 1/4$ has an order of magnitude given by T^{-1}, while the corresponding order for the bias is T^{2d-1}, uniformly in i. Noticing that $k/T \to \infty$, it follows that (55) tends to zero, which implies the desired result.

The uniform convergence follows from the pointwise convergence and the same arguments as in Lemma 1.

Lemma 4 *Let $\hat{J}_{ij}(\lambda_0)$ be the (i,j) element of the $k \times (p+q+1)$ Jacobian matrix $\hat{J}_k(\lambda_0)$ and $J_{ij}(\lambda_0)$ be its probability limit defined in (47). Under the assumptions of Sections 2 and 3 it holds that:*

1. *$sup_{1\leq i\leq k}\left|\hat{J}_{ij}(\lambda_0) - J_{ij}(\lambda_0)\right| \xrightarrow{p} 0$, for all $j = 1, ..., p+q+1$.*

2. *$lim_{T,k\to\infty}\hat{J}_k(\lambda_0)\hat{J}_k(\lambda_0) = \lim_{k\to\infty}\Xi_k = \Xi$, defined as,*

$$\Xi = \begin{pmatrix} \pi^2/6 & \Pi' \\ \Pi & \Xi_{pq} \end{pmatrix}$$

where $\Pi = (\pi_\omega(0), ..., \pi_\omega(p-1), \pi_\psi(0), ..., \pi_\psi(q-1))$,

$\pi_\omega(j) = \sum_{i=0}^{\infty}\frac{\omega_i}{j+i+1}$, $\pi_\psi(j) = \sum_{i=0}^{\infty}\frac{\psi_i}{j+i+1}$ and Ξ_{pq} is the Fisher information matrix corresponding to pure ARMA processes. Then, Ξ coincides with the Fisher information matrix for ARFIMA processes (see, for instance, Li and McLeod, 1986).

Proof of Lemma 4

1. Let $j = 1$. The first column of $\hat{J}(\lambda_0)$ contains the derivatives of the autocorrelation function with respect to d, i.e., $\hat{J}_{i1} = \partial\hat{\rho}_{e(\lambda_0)}(i)/\partial d$ with $i = 1, ...k$, that are given by,

$$\frac{\partial\hat{\rho}_{e(\lambda_0)}(i)}{\partial d} = \frac{\partial\tilde{\rho}_{e(\lambda_0)}(i)}{\partial d}\frac{T\sigma^2}{\sum e_t^2(\lambda_0)} + \tilde{\rho}_{e(\lambda_0)}(i)\frac{T\sigma^2\sum e_t(\lambda_0)e_t'(\lambda_0)}{\left(\sum e_t^2(\lambda_0)\right)^2},$$

where $\tilde{\rho}_{e(\lambda_0)}(i)$ is defined as in (35). The elements of the first column of the matrix $J(\lambda_0)$ are given by $J_{i1} = -1/i$, $i = 1, ...k$, (see (47)). Then,

$$sup_{1\leq i\leq k}\left|\frac{\partial\hat{\rho}_{e(\lambda_0)}(i)}{\partial d} + \frac{1}{i}\right| \leq sup_{1\leq i\leq k}\left|\frac{\partial\tilde{\rho}_{e(\lambda_0)}(i)}{\partial d} + \frac{1}{i}\right| \tag{56}$$

$$+sup_{1\leq i\leq k}\left|\tilde{\rho}_{e(\lambda_0)}(i)\right|\frac{T\sigma^2\sum e_t(\lambda_0)e_t'(\lambda_0)}{\left(\sum e_t^2(\lambda_0)\right)^2} + o_p(1), \tag{57}$$

since $\frac{T\sigma^2}{\sum e_t^2(\lambda_0)} = 1 + o_p(1)$ and does not depend on i.

We will make use of the following fact (see Bai and Ng, (2004, p. 1155): Let $X_1, ..., X_k$ be an arbitrary sequence of random variables. If $\max_{1\leq i\leq k} E|X_k|^\alpha < M$, with $\alpha > 0$, then $\max_{1\leq i\leq k}|X_k| = O_p\left(k^{1/\alpha}\right)$. Let $X_i = T^{1/2}\left(\frac{\partial\hat{\rho}_{e(\lambda_0)}(i)}{\partial d} + \frac{1}{i}\right)$ and consider

$$\max_{1\leq i\leq k} E\left(T^{1/2}\left(\frac{\partial\tilde{\rho}_{e(\lambda_0)}(i)}{\partial d} + \frac{1}{i}\right)\right)^2 \tag{58}$$

$$
\begin{aligned}
&= \max_{1\leq i\leq k} TE\left(\frac{\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}+\sum\log(1-L)\,\varepsilon_t\varepsilon_{t+i}}{T\sigma^2}+\frac{1}{i}\right)^2 \\
&= \max_{1\leq i\leq k}\left\{E\left(\frac{\left(\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}\right)^2}{T\sigma^4}\right)+E\left(\frac{\left(\sum_{t=1}^{T-i}\varepsilon_t\log(1-L)\,\varepsilon_{t+i}\right)^2}{T\sigma^4}\right)\right. \\
&\quad +TE\left(\frac{2\left(\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}\right)\left(\sum_{s=1}^{T-i}\log(1-L)\,\varepsilon_{s+i}\varepsilon_s\right)}{T^2\sigma^4}\right) \\
&\quad \left.+TE\left(2\frac{1}{i}\frac{\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}+\sum\log(1-L)\,\varepsilon_t\varepsilon_{t+i}}{T\sigma^2}\right)+\frac{T}{i^{-2}}\right\}.
\end{aligned}
$$

Using the expansion $\log(1-L) = -(L+L/2+L/3+...)$, it is easy to get:

$$
E\left(\frac{\left(\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}\right)^2}{T\sigma^4}\right) = \left(T^{-1}\sum_{t=1}^{T-i}\left(\sum_{l=1}^{t-1}l^{-2}\right)\right), \tag{59}
$$

and

$$
E\left(\frac{\left(\sum_{t=1}^{T-i}\varepsilon_t\log(1-L)\,\varepsilon_{t+i}\right)^2}{T\sigma^4}\right) \leq \left\{T^{-1}\sum_{t=1}^{T-i}\left(\sum_{l=1}^{t+i-1}l^{-2}\right)+\frac{\mu_4(T-i)+(T-i)^2}{Ti^2}\right\}. \tag{60}
$$

With respect to the cross products:

$$
E\left(2\frac{1}{i}\frac{\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}+\sum\log(1-L)\,\varepsilon_t\varepsilon_{t+i}}{T\sigma^2}\right) = -\frac{2(T-i)}{i^2},
$$

and

$$
E\left(\frac{\left(\sum_{t=1}^{T-i}\log(1-L)\,\varepsilon_t\varepsilon_{t+i}\right)\left(\sum_{s=1}^{T-i}\log(1-L)\,\varepsilon_{s+i}\varepsilon_s\right)}{T\sigma^4}\right) = 0.
$$

It follows from the results above that (58) is bounded by,

$$
\begin{aligned}
&\max_{1\leq i\leq k}\left\{2T^{-1}\sum_{t=1}^{T-i}\left(\sum_{l=1}^{t+i-1}l^{-2}\right)+\frac{\mu_4(T-i)}{Ti^2}+\frac{(T-i)^2-2(T-i)T+T^2}{Ti^2}\right\} \\
&= \max_{1\leq i\leq k}\left\{2T^{-1}\sum_{t=1}^{T-i}\left(\sum_{l=1}^{t+i-1}l^{-2}\right)+\frac{\mu_4(T-i)}{Ti^2}+\frac{1}{T}\right\} \leq 2T^{-1}\sum_{t=1}^{T}\left(\sum_{l=1}^{t+k-1}l^{-2}\right)+\mu_4+\frac{1}{T} < (61)
\end{aligned}
$$

where $M' < \infty$ is a constant. This implies that $\max_{1\leq i\leq k}\left|T^{1/2}\left(\frac{\partial\tilde{\rho}_{e(\lambda_0)}(i)}{\partial d}+\frac{1}{i}\right)\right| = O_p\left(k^{1/2}\right)$ and consequently,

$$\max_{1\leq i\leq k}\left|\left(\frac{\partial\tilde{\rho}_{e(\lambda_0)}(i)}{\partial d}+\frac{1}{i}\right)\right| = O_p\left(k^{1/2}/T^{1/2}\right) = o_p(1) \tag{62}$$

since $k/T \to 0$. Following the same steps it is easy to check that $sup_{1\leq i\leq k}\left|\tilde{\rho}_{e(\lambda_0)}(i)\right| \to 0$ and since $\frac{T\sigma^2\sum e_t(\lambda_0)e_t'(\lambda_0)}{\left(\sum e_t^2(\lambda_0)\right)^2} = o_p(1)$, it follows that the LHS of (56) tends to zero as desired.

The proof for the remaining columns of the matrix $\hat{J}(\lambda_0)$ is analogous and therefore it is omitted.

2. It follows from Lemma 2 that for any fixed k, $\lim_{T\to\infty}\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0) = J_k'(\lambda_0)J_k(\lambda_0)$. On the other hand, it is straight forward to check that $\lim_{k\to\infty}J_k'(\lambda_0)J_k(\lambda_0) = \Xi$, since this limit just involves non-stochastic terms. Then (see Lemma 6 in Phillips and Moon, 1999), if,

$$\lim\sup_{k,T\to\infty} P\left(\left\|\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0) - J_k'(\lambda_0)J_k(\lambda_0)\right\| > \epsilon\right) = 0 \text{ for all } \epsilon > 0 \tag{63}$$

holds, then the sequential and the joint limit coincide, that is, $\lim_{T,k\to\infty}\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0) = \Xi$.

By Markov's inequality, the probability in (63) is bounded by:

$$\epsilon^{-1}E\left\|\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0) - J_k'(\lambda_0)J_k(\lambda_0)\right\|. \tag{64}$$

For simplicity, consider first the case where $\lambda = d$. In this case the expectation in (64) is given by,

$$\epsilon^{-1}E\left|\sum_{i=1}^{k}\left(\frac{\partial\hat{\rho}_{e(d_0)}^2(i)}{\partial d} - i^{-2}\right)\right| = \epsilon^{-1}\left|\sum_{i=1}^{k}E\left(\frac{\partial\hat{\rho}_{e(d_0)}^2}{\partial d}(i)\right) - \sum_{i=1}^{k}i^{-2}\right|. \tag{65}$$

Taking into account that,

$$\sum_{i=1}^{k}E\left(\frac{\partial\hat{\rho}_{e(d_0)}^2}{\partial d}(i)\right) = T^{-2}\sum_{i=1}^{k}\left(\sum_{t=1}^{T-i}\sum_{s=1}^{t-1}s^{-2} + \sum_{t=1}^{T-i}\sum_{s=1}^{t+i-1}s^{-2} + \mu_4(T-i) + (T-i)^2\frac{1}{i^{-2}}\right) + o(1) \tag{66}$$

and the fact that $k/T \to 0$, it follows that all the terms in the right hand side of (66) tend to zero but for

$$T^{-2}\sum_{i=1}^{k}\frac{(T-i)^2}{i^{-2}}$$

which tends to $\pi^2/6$ when both k and T tend to infinity. This implies that (65) also tends to zero, since $\lim_{k\to\infty}\sum_{i=1}^{k} i^{-2} = \pi^2/6$.

The general case where other ARMA parameters are estimated can be solved along the same lines.■

Proof of Theorem 5.

The proof of the first part of the theorem is analogous to that of Theorem 3, with Definition 1 and Lemma 1 replaced by Definition 2 and the results of Lemma 3.

First step. Define $\hat{\lambda}_1 = \arg\min_{\lambda\in\Lambda_1^*} V_{ke}(\lambda)$. Using the same arguments as in Theorem 3, it follows that $\hat{\lambda}_1 \xrightarrow{p} \lambda_0$.

Second step. Consider the case where $d_0 > 1/4 + \nabla_1$ and define $\Lambda_2^* = \{d : \nabla_1 \leq d < d_0 - 1/4 + \eta\} \times \Lambda^{(-1)}$ and $\hat{\lambda} = \arg\min_{\lambda\in\Lambda} V_{ke}(\lambda)$. It is needed to show that for any $\delta > 0$, $P\left(\left\|\hat{\lambda} - \hat{\lambda}_1\right\| \geq \delta\right) \to 0$. This probability is bounded by $P\left(\inf_{\lambda\in\Lambda_2^*} V_{ke}(\lambda) \leq 0\right) + P\left(V_{ke}\left(\hat{\lambda}_1\right) > 0\right)$ which tends to zero by the same arguments as in Theorem 3.

With respect to the asymptotic distribution of $\hat{\lambda}$, application of the mean value theorem to the first-order conditions yields an expression similar to (52), where $\bar{\lambda}$ can be replaced by λ_0 by the same reasons as in Theorem 4. By Theorem 2 and Lemma 4 it follows that $\left(\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0)\right)^{-1}\hat{J}_k'(\lambda_0)\sqrt{T}\hat{\rho}_{ke(\lambda_0)} \to N(0, \Xi^{-1})$, as desired.■

Proof of Theorem 6

Let $\hat{\rho}_{ke(\hat{\lambda})}$ be the vector defined in (21). A first-order Taylor's series expansion of $\hat{\rho}_{ke(\hat{\lambda})}$ around λ_0 yields,

$$\hat{\rho}_{ke(\hat{\lambda})} = \hat{\rho}_{ke(\lambda_0)} + \frac{\partial\hat{\rho}_{ke(\lambda^*)}}{\partial\lambda'}\left(\hat{\lambda} - \lambda_0\right). \tag{67}$$

By Theorem 1, $\sqrt{T}\hat{\rho}_{ke(\lambda_0)} = \sqrt{T}\hat{\rho}_{k\varepsilon} + o_p(1) \xrightarrow{w} N(0, I_k)$ for $d > -0.75$. Since the derivative of $\frac{\partial\hat{\rho}_{ke(\lambda^*)}}{\partial\lambda'}$ exists and is bounded in an open convex set containing λ_0 and $\lambda^* \xrightarrow{p} \lambda_0$, then it is possible to substitute λ^* by λ_0 in (67). The joint distribution of $\left(\hat{J}_k(\lambda_0)\left(\hat{\lambda} - \lambda_0\right), \hat{\rho}_{k\varepsilon}\right)'$ is given by:

$$\sqrt{T}\begin{pmatrix} \hat{J}_k(\lambda_0)\left(\hat{\lambda} - \lambda_0\right) \\ \hat{\rho}_{k\varepsilon} \end{pmatrix} = \begin{pmatrix} -\hat{J}_k(\lambda_0)\left(\hat{J}_k'(\lambda_0)\hat{J}_k(\lambda_0)\right)^{-1}\hat{J}_k'(\lambda_0) \\ I_k \end{pmatrix}\sqrt{T}\hat{\rho}_{k\varepsilon} + o_p(1)$$

$$\xrightarrow{w} N(0, \Upsilon),$$

where:

$$\Upsilon = \begin{pmatrix} -J_k(\lambda_0)\left(J_k'(\lambda_0)J_k(\lambda_0)\right)^{-1}J_{\lambda_0 k}' & J_k(\lambda_0)\left(J_k'(\lambda_0)J_k(\lambda_0)\right)^{-1}J_k'(\lambda_0) \\ J_k(\lambda_0)\left(J_k'(\lambda_0)J_k(\lambda_0)\right)^{-1}J_k'(\lambda_0) & I_k \end{pmatrix}.$$

Since joint normality holds, any linear combination of $\left(\hat{J}_k(\lambda_0)\left(\hat{\lambda}-\lambda_0\right), \hat{\rho}_{k\varepsilon}\right)$ will also be normal. Taking into account expression (67) it follows that:

$$\sqrt{T}\hat{\rho}_{ke(\hat{\lambda})} \xrightarrow{w} N\left(0, I_k - J_k(\lambda_0)\left(J_k'(\lambda_0)J_k(\lambda_0)\right)^{-1}J_k'(\lambda_0)\right). \blacksquare$$

Appendix 2

This appendix reports the results of some Monte-Carlo experiments not included in the main text. Table 12 presents the mean and standard deviation of correlations at different lags associated to the (truncated) residual process $e_t(d_0)$ for different values of d_0. To obtain these residuals, the following procedure has been implemented: processes of the form $\Delta^{\varphi_0} y_t = \varepsilon_t$ were generated for a large sample size equal to $n+T$ and then the first $n = 1000$ observations were rejected. The last T observations were integrated an integer number of times, m_0. Truncated residuals were computed by applying the finite filter $\sum_{i=0}^{t-1} \pi_i L^i$ to the process y_t, where the coefficients $\{\pi_i\}$ come from the expansion in powers of L of the polynomial $(1-L)^{d_0}$. Again the results in finite samples confirm the asymptotic results. It can be seen that for invertible processes ($d_0 > -1$), estimated correlations behave correctly in the sense that, as expected, zero mean and unit variance is found. Nevertheless, for non-invertible processes, residuals correlations do not provide consistent estimates of the innovation correlations.

Table 12. Residual Autocorrelations

d_0	-1.2	-1.0	-0.7	0.4	0.8	1.4
$T = 100$						
mean($\sqrt{T}\hat{\rho}_e(1)$)	0.801	-0.136	-0.063	-0.061	-0.128	-0.068
std($\sqrt{T}\hat{\rho}_e(1)$)	1.435	0.993	0.964	0.966	0.955	0.979
mean($\sqrt{T}\hat{\rho}_e(2)$)	0.804	-0.085	-0.022	-0.037	-0.080	-0.027
std($\sqrt{T}\hat{\rho}_e(2)$)	1.386	0.971	0.977	0.979	0.995	0.990
mean($\sqrt{T}\hat{\rho}_e(5)$)	0.793	-0.046	0.009	-0.006	-0.042	-0.002
std($\sqrt{T}\hat{\rho}_e(5)$)	1.361	0.956	0.966	0.964	0.971	0.974
$T = 400$						
mean($\sqrt{T}\hat{\rho}_e(1)$	3.020	-0.006	0.010	0.047	-0.001	0.042
std($\sqrt{T}\hat{\rho}_e(1)$	3.248	0.979	1.001	0.982	0.979	0.983
mean($\sqrt{T}\hat{\rho}_e(2)$)	2.926	-0.045	-0.031	-0.005	-0.039	-0.005
std($\sqrt{T}\hat{\rho}_e(2)$)	3.170	1.013	1.027	1.016	1.013	1.019
mean($\sqrt{T}\hat{\rho}_e(5)$)	2.915	-0.006	0.002	0.026	-0.002	0.029
std($\sqrt{T}\hat{\rho}_e(5)$)	3.149	0.968	1.031	0.958	0.969	0.964