Minimum distance estimation of stationary and non-stationary
ARFIMA processes

LAURA MAYORAL

Institute for Economic Analysis (CSIC) and Universidad Pompeu Fabra
E-mail: laura.mayoral@upf.edu

Received: January 2006

Summary 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 and asymptotically normally distributed 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.

Key words: Fractional integration, Nonstationarity, Long-memory, Minimum distance estimation.

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 of providing a good characterization of the long-run properties of many economic and financial time series. They are also very useful for modelling multivariate time series, since they can capture a larger number of long-term equilibrium relations among economic variables than the traditional multivariate ARIMA models. See Henry and Zaffaroni (2002) and Robinson (2003) for recent 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 estimators considered in Tieslau, Schmidt and Baillie (1996), Chung and Schmidt (1995) and Chong (2000). Nevertheless, as it will be seen shortly, it presents important advantages over those

© Royal Economic Society 2007. Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA, 02148, USA.
estimators. It is denoted by ‘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 very 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 and that it is computationally faster than exact MLE for more complex ARFIMA $(p, d, q)$ processes. As a drawback, the proposed estimator is not asymptotically efficient as MLE is, although it will be shown that the asymptotic variance of the former estimator can be arbitrarily close to the inverse of the Fisher information matrix and therefore, very close to efficiency for all practical purposes.

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; 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 the 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.

It is also remarkable that the proposed framework can be easily extended to more general settings. For instance, the estimator can be easily robustified against conditional heterocedasticity, just by introducing a modification in the definition of the residuals used to compute the autocorrelations. Finally, it can be extended, along the lines of Wright (1999), to the estimation of the fractionally integrated stochastic volatility model.

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 and Pierce (1970) and Box and Ljung (1978) goodness-of-fit statistics while 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 the Appendix.

The following conventional notation is adopted throughout the paper. $L$ is the lag operator; $\Delta = (1 - L)$; $\Gamma(.)$ denotes the gamma function; $\overset{w}{\rightarrow}$ and $\overset{P}{\rightarrow}$ denote weak convergence and convergence in probability, respectively; $\{\pi_i(d)\}_{i=0}^{\infty}$ represents the sequence of coefficients associated with the expansion of $\Delta^d$ in powers of $L$, such that $\Delta^d = \pi_0(d) + \pi_1(d)L + \pi_2(d)L^2$.
$L^2 + \ldots$, and

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

2. THE MODEL

The process $y_t$, observed at time $t = 1, \ldots T$, 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,

$$\Phi_0(L) \Delta^{d_0} (y_t - \mu_0) = \Theta_0(L) \epsilon_t, \ t = 0, \pm 1, \ldots,$$  \hspace{1cm} (2)

where $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is a sequence of i.i.d. zero-mean random variables with unknown variance $\sigma^2$ and finite fourth moment, $E(\epsilon_t^4) = \mu_4 < \infty$. $\Phi_0(L)$ 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.\(^1\) For non-stationary values of $d_0$($d_0 \geq 0.5$), we assume that the process $y_t$ begins at time $t = 1$, that is,\(^2\)

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

where

$$\Phi_0(L) \Delta^{m_0} (x_t(m_0) - \mu_0) = \Theta_0(L) \epsilon_t, \ t = 0, \pm 1, \ldots$$  \hspace{1cm} (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 $\lfloor \cdot \rfloor$ denotes integer part, is the number of times that $y_t$ must be differenced to achieve stationarity (therefore $m_0 \geq 0$). The parameter $\varphi_0$, 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 $\varphi_0$. For $m_0 = 0$, $\mu_0$ 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.

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, $\mu_0$, is known or unknown.

---

\(^1\)This assumption, although restrictive, is standard when parametric estimation is considered. Several authors have assessed the performance of model selection criteria for ARFIMA processes (see Beran, Bhwsali and Ocker 1998; Smith, Taylor and Yadav 1997; Crato and Rey 1996). It is concluded that the BIC usually outperforms other standard selection criteria.

\(^2\)See Marinucci and Robinson (1999) for the different definitions of non-stationary ARFIMA processes and their asymptotic implications.
2.1. Residuals when $\mu_0$ is known and equal to zero

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

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

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 $\Phi_0(L)\Theta_0(L)^{-1}\Delta^{\psi_0}$ in powers of $L$. Given the observations $y_1, \ldots, y_T$, the innovations $\varepsilon_t$ cannot be computed directly, since an infinite sample would be needed. Nevertheless, they may be estimated by

$$e_t(\lambda) = \sum_{j=0}^{t-m-1} \alpha_j(\lambda) x(m_{t-j}), \quad t = m + 1, \ldots, T.$$

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

2.2. Residuals when $\mu_0$ is unknown

When $\mu_0$ is unknown, the residuals defined above need to be adjusted. Again, following Beran (1995), we define

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

Note that since $x_t(m_0)$ is stationary and ergodic, the sample mean $\bar{x}(m_0)$ is a consistent estimator of $\mu_0$. For some integer and positive $m$, adjusted residuals are given by

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

where $\{\alpha_j(\lambda_0)\}_{j=0}^{\infty}$ are the coefficients associated with the expansion of $\Phi(L)\Theta(L)^{-1}\Delta^\psi$ in powers of $L$.

More efficient estimators of $\mu_0$ have been proposed by Adenstedt (1974) and could also be used to compute the estimator presented in Section 3. Nevertheless, since these estimators depend on the remaining unknown parameters, the computation would become more involved. Since the

---

3This 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.

4Note that to compute the adjusted residuals $m$ is not assumed to be known. The estimation procedure chooses the value of $\hat{d}$ (hence, of $\hat{m}$ and $\hat{\varphi}$) that minimizes the criterion function in (13).

© Royal Economic Society 2007
efficiency improvement in the estimation of $\mu_0$ does not affect the asymptotic properties of the estimates of the remaining parameters, we advocate the use of the sample mean for estimating $\mu_0$.\(^5\)

3. GENERALIZED MINIMUM DISTANCE ESTIMATION OF ARFIMA PROCESSES

Minimum Distance (MD) is a classical estimation approach in the econometric literature. This technique encompasses other popular procedures such as Generalized Method of Moments (GMM), Non-Linear Least Squares (NLS) or Maximum Likelihood (ML) among others. MD estimators are computed by minimizing the 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, $\lambda$, such that it verifies that $\hat{g}_T(\lambda_0) \xrightarrow{p} 0$, where $\lambda_0$ is the true parameter value; $\hat{W}$ is a positive definite weighting matrix that defines the distance. Under standard regularity conditions, the resulting estimator is $\sqrt{T}$-consistent and asymptotically normally distributed (see Newey and McFadden 1994). Different choices of the function $\hat{g}_T(\lambda)$ 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. With respect to the choice of $\hat{W}$, it is well-known that if $\text{var}(\sqrt{T} \hat{g}_T(\lambda_0)) \xrightarrow{p} \Omega$, 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 $(J_{\lambda_0}^{'} \Omega^{-1} J_{\lambda_0})^{-1}$, where $J_{\lambda_0}$ is the limit of the Jacobian matrix of $\hat{g}_T$. Hence, this would be in general a good choice. As it becomes clear from the discussion above, one of the main advantages of MD relative to ML estimation is that the former does not require to assume a particular distribution of the innovation sequence at any stage.

Several parametric MD techniques can be found in the literature of FI processes. Of particular relevance to this paper is the MD estimator proposed by Tieslau et al. (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:

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

where $\hat{\rho}_{k}$ is the sample autocorrelation function of the (stationary) process $y_t$ up to lag $k$ (for a fixed value of $k$), $\rho_{k} (\lambda)$ is the theoretical autocorrelation of the corresponding ARFIMA $(p, d, q)$ process up to the same lag and $\hat{W}$ is a symmetric, positive definite weighting matrix. Chong (2000) proposed a related estimator that uses partial correlations instead of correlations to form the moment conditions. Although theoretically very appealing, there remain significant problems with both procedures. First, they are restricted to stationary series since they require the existence of autocorrelations. And second, they are $\sqrt{T}$-consistent and asymptotically normal only for $d_0 < 0.25$, due to the non-standard behaviour of sample autocorrelations of ARFIMA processes outside this range. Chung and Schmidt (1995) have demonstrated, by applying the results on

\(^5\)Alternatively, the asymptotically efficient estimator of $\mu_0$ proposed by Adenstedt (1974) can be computed in a second step, after estimating the remaining parameters with the method proposed in this paper, since knowledge of $d_0$ is required.

© Royal Economic Society 2007
Minimum distance estimation

129

correlations 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 (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 non-stationary series, since the existence of autocorrelations is not required, but its consistency has not been proved yet in this general framework.

Let us now describe the GMD estimator proposed in this paper. Consider the sample \( i \)-th autocorrelation associated with the residuals defined in (5) or in (7) given by

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

Also define the vector \( \hat{\rho}_{ke(\lambda)} \) that contains the first \( k \) sample autocorrelations,

\[
\hat{\rho}_{ke(\lambda)} = (\hat{\rho}_{e(\lambda)} (1), \ldots, \hat{\rho}_{e(\lambda)} (k))'.
\] (11)

The following theorem will be very useful in the derivation of the subsequent theory.

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

\[
\sqrt{T} (\hat{\rho}_{ke(\lambda_0)} - \hat{\rho}_{ke}) = o_p (1),
\] (12)

where \( \hat{\rho}_{ke} = (\hat{\rho}_{e}(1), \ldots, \hat{\rho}_{e}(k))' \) is the vector that contains the first \( k \) sample autocorrelations of the innovations \( \varepsilon_t \).

The process \( \sqrt{T} \hat{\rho}_{ke} \) converges as \( T \rightarrow \infty \) to a zero-mean Gaussian distribution with variance-covariance matrix equal to the identity matrix (see theorem 7.2.1 and example 7.2.1 in Brockwell and Davis (1991)). In turn, Theorem 1 ensures that the asymptotic distribution of \( \hat{\rho}_{ke(\lambda_0)} \) and \( \hat{\rho}_{ke} \) is the same. Hence,

\[
\sqrt{T} \hat{\rho}_{ke(\lambda_0)} \overset{w}{\rightarrow} N (0, I_k),
\]

where \( I_k \) denotes the identity matrix of order \( k \). This result implies that the effect of the truncation that has to be introduced to compute the residuals \( e_t(\lambda_0) \) has a negligible asymptotic impact in the distribution of the sample autocorrelations.

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 \( I_k \), (Theorem 1), it follows that the efficient weighting matrix is \( I_k \). Moreover, since \( \hat{\rho}_{ke(\lambda_0)} \rightarrow 0 \), the MD criterion function, \( V_{ke} \), becomes

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

© Royal Economic Society 2007
and the GMD estimator \( \hat{\lambda}_k \) is defined as
\[
\hat{\lambda}_k = \arg \min_{\lambda \in \Lambda} V_{ke}(\lambda, y).
\] (14)

Note that, since all the sample autocorrelations converge to zero 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. Furthermore, only the existence of residual autocorrelations is required but not those of the original series. Hence, it can also be applied to non-stationary series, in contrast to the TSB or the Chong estimator.

Theorems 2 and 3 describe the asymptotic properties of the GMD estimator proposed here. The former theorem states its consistency, while the latter presents its asymptotic distribution.

**Theorem 2** Let \( y_t \) be an ARFIMA\((p, d, q)\) process under the hypotheses of Section 2. Also let \( \lambda_0 \) be an interior point of the compact set \( \Lambda \). Then, as \( T \) tends to infinity, it holds that
\[
\hat{\lambda}_k \overset{P}{\to} \lambda_0,
\]
where \( \hat{\lambda}_k \) is the GMD estimator defined in (14) and \( k \) is a fixed number such that \( p + q + 1 \leq k \leq T - 1 \).

**Theorem 3** Under the hypotheses of the previous theorem, it holds that
\[
\sqrt{T}(\hat{\lambda}_k - \lambda_0) \xrightarrow{w} N(0, \Xi_k^{-1}),
\] (15)
where \( \Xi_k = J'_k(\lambda_0) J_k(\lambda_0) \) and
\[
J_k(\lambda_0) = \begin{pmatrix}
-1 & 1 & 0 & \ldots & 1 & \ldots & 0 \\
-1/2 & \omega_1 & 1 & \ldots & \psi_1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-1/k & \omega_{k-1} & \omega_{k-2} & \ldots & \psi_{k-1} & \ldots & \psi_{k-q}
\end{pmatrix}.
\] (16)

(See the Appendix for the definition of the coefficients \( \omega(\cdot) \) and \( \psi(\cdot) \)).

The estimator defined in (14) is not efficient since the matrix \( \Xi_k \) is not the Fisher information matrix. Nevertheless, it is easy to check that
\[
\lim_{k \to \infty} \Xi_k = \Xi,
\] (17)
where \( \Xi \) is the Fisher information matrix for ARFIMA processes (see Fox and Taqqu, 1986) defined as
\[
\Xi = \begin{pmatrix}
\pi^2/6 & \Pi' \\
\Pi & \Xi_{pq}
\end{pmatrix},
\] (18)
where \( \Pi = (\pi_\omega(0), \ldots, \pi_\omega(p - 1), \pi_\psi(0), \ldots, \pi_\psi(q - 1)) \) and
\[
\pi_\omega(j) = \sum_{i=0}^{\infty} \frac{\omega_i}{j + i + 1}, \quad \pi_\psi(j) = \sum_{i=0}^{\infty} \frac{\psi_i}{j + i + 1}.
\]

The result in (17) means that for \( T \) large enough, it is possible to select a value \( k^* \) for which the matrices \( \Xi \) and \( \Xi_{k^*} \) are arbitrarily close, which implies that the corresponding estimator \( \hat{\lambda}_{k^*} \)
Minimum distance estimation

would be very close to be efficient. To see this more clearly, consider the ARFIMA \((0, d, 0)\) case. The asymptotic lower bound for the standard deviation of any estimate of \(d\) is given by \(T^{-1/2}(\pi^2/6)^{-1/2}\) while the standard deviation of \(\hat{\lambda}_k\) is given by \(T^{-1/2}(\sum_{j=1}^{k} \frac{1}{j^2})^{-1/2}\). This implies that the relative efficiency of the latter with respect to the former is 94.5%, 97.1%, 98.5% and 99.0% for \(k = 5, 10, 20\) and 30, respectively. Then, moderate values of \(k\) are able to provide very good approximations to the Cramer–Rao lower bound.

Asymptotically, the larger the value of \(k\) chosen to compute \(\hat{\lambda}_k\), the higher the efficiency. But interestingly, this result is not true in finite samples. The next section analyzes the issue of selecting the value of \(k\) for a given sample size via Monte Carlo simulation.

Finally, it is interesting to note that many estimation approaches developed for ARFIMA processes are only valid in the invertible and stationary range of values of \(d\), \((-0.5 < d < 0.5)\). This implies that the researcher should first determine \(m_0\) in an exploratory way. If the latter value is incorrectly guessed, the former family of estimators becomes inconsistent. For this reason, it is always safer to use a method that covers the whole range of values of \(d\).

4. BEHAVIOUR OF THE GMD ESTIMATOR IN FINITE SAMPLES

In this section, a Monte Carlo study is conducted to investigate the finite-sample performance of the MD estimator defined in (14). Processes of the form \(\Delta^q(\Delta^{m_0} y_t - \mu_0) = u_t\) have been generated, with different specifications for \(u_t\) that will be detailed below. Innovations have been drawn from independent \(N(0, 1)\) distributions unless otherwise noted. Further, \(\mu_0\) is set equal to zero and computations have been carried out considering that this value is both known and unknown.

Prior to estimating the above-mentioned models, it is necessary to select the number of correlations, \(k\), to be included in the criterion function \(V_{ke}\). Although increasing \(k\) always improves efficiency asymptotically, the suitable choice of \(k\) in finite-sample applications depends on the sample size, the number of parameters to be estimated \((p + q + 1)\) and the values of the 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 through Monte Carlo simulation. Hence, for a given sample size \(T\) the value of \(k\) has been chosen as a function of that quantity. In particular, \(k\) is set equal to 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). Next, the performance of each of these values has been examined for different values of \(T\) and across several models that differ on the number of parameters to be estimated.

Table 1 presents the bias and the square root of the mean square error (\(\sqrt{MSE}\)) of the GMD estimator for two different sample sizes, \(T = 100\) and 400 and different values of \(k\) for the case where \(u_t = \varepsilon_t\) and \(\mu_0\) is known. Table 2 presents analogous results for the case where \(\mu_0\) is unknown and has to be estimated. The number of replications is 5000. The GMD estimator is surprisingly robust across the values of \(k\). Moderate values of \(k\) tend to provide the best results.

6It would be possible to prove a joint asymptotic result, with \(T\) and \(k\) tending to infinity simultaneously, following Phillips and Moon (1999).

7The asymptotic theory in Theorems 2 and 3 has been derived for the case where \(k\) is fixed and not an increasing function of \(T\). Nevertheless, note that the experiments provided in this section are designed to facilitate the choice of \(k\) in the finite-sample case where \(T\) is also fixed. Therefore there is not a contradiction between both set of results.
<table>
<thead>
<tr>
<th>$d_0$</th>
<th>-0.7</th>
<th>-0.2</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
<th>1.4</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = T^{1/4}$</td>
<td>Bias $\hat{d}$</td>
<td>-0.025</td>
<td>-0.020</td>
<td>-0.018</td>
<td>-0.016</td>
<td>-0.021</td>
<td>-0.017</td>
<td>-0.024</td>
<td>-0.019</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/3}$</td>
<td>-0.020</td>
<td>-0.021</td>
<td>-0.020</td>
<td>-0.025</td>
<td>-0.022</td>
<td>-0.029</td>
<td>-0.035</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/2}$</td>
<td>-0.022</td>
<td>-0.023</td>
<td>-0.024</td>
<td>-0.027</td>
<td>-0.024</td>
<td>-0.032</td>
<td>-0.036</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/4}$</td>
<td>0.101</td>
<td>0.092</td>
<td>0.090</td>
<td>0.097</td>
<td>0.094</td>
<td>0.096</td>
<td>0.097</td>
<td>0.093</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{d}$</td>
<td>$k = T^{1/3}$</td>
<td>0.111</td>
<td>0.110</td>
<td>0.101</td>
<td>0.112</td>
<td>0.111</td>
<td>0.109</td>
<td>0.114</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/2}$</td>
<td>0.116</td>
<td>0.117</td>
<td>0.115</td>
<td>0.115</td>
<td>0.114</td>
<td>0.113</td>
<td>0.122</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/4}$</td>
<td>-0.003</td>
<td>-0.065</td>
<td>-0.007</td>
<td>-0.008</td>
<td>-0.008</td>
<td>-0.010</td>
<td>-0.008</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/3}$</td>
<td>-0.003</td>
<td>-0.009</td>
<td>-0.008</td>
<td>-0.009</td>
<td>-0.010</td>
<td>-0.012</td>
<td>0.008</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/2}$</td>
<td>-0.003</td>
<td>-0.009</td>
<td>-0.010</td>
<td>-0.011</td>
<td>-0.010</td>
<td>-0.013</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/4}$</td>
<td>0.045</td>
<td>0.041</td>
<td>0.041</td>
<td>0.042</td>
<td>0.045</td>
<td>0.041</td>
<td>0.042</td>
<td>0.046</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{d}$</td>
<td>$k = T^{1/3}$</td>
<td>0.044</td>
<td>0.040</td>
<td>0.041</td>
<td>0.043</td>
<td>0.043</td>
<td>0.041</td>
<td>0.042</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>$k = T^{1/2}$</td>
<td>0.047</td>
<td>0.042</td>
<td>0.042</td>
<td>0.045</td>
<td>0.045</td>
<td>0.042</td>
<td>0.044</td>
<td>0.048</td>
</tr>
</tbody>
</table>
Table 2. Estimation of $d$ DGP: ARFIMA (0, $d$, 0); $\mu_0$ unknown ($\mu_0 = 0$).

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>$-0.7$</th>
<th>$-0.2$</th>
<th>$0$</th>
<th>$0.2$</th>
<th>$0.4$</th>
<th>$0.8$</th>
<th>$1.0$</th>
<th>$1.4$</th>
<th>$2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = T^{1/4}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias $\hat{d}$</td>
<td>$0.101$</td>
<td>$0.106$</td>
<td>$0.101$</td>
<td>$0.099$</td>
<td>$0.097$</td>
<td>$0.103$</td>
<td>$0.104$</td>
<td>$0.100$</td>
<td>$0.104$</td>
</tr>
<tr>
<td>$\sqrt{MSE} \hat{d}$</td>
<td>$0.113$</td>
<td>$0.108$</td>
<td>$0.110$</td>
<td>$0.0105$</td>
<td>$0.106$</td>
<td>$0.109$</td>
<td>$0.110$</td>
<td>$0.114$</td>
<td>$0.107$</td>
</tr>
<tr>
<td>$k = T^{1/2}$</td>
<td>$0.117$</td>
<td>$0.109$</td>
<td>$0.108$</td>
<td>$0.111$</td>
<td>$0.111$</td>
<td>$0.112$</td>
<td>$0.123$</td>
<td>$0.117$</td>
<td>$0.117$</td>
</tr>
<tr>
<td>$k = T^{1/3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias $\hat{d}$</td>
<td>$0.009$</td>
<td>$0.009$</td>
<td>$0.003$</td>
<td>$0.001$</td>
<td>$0.003$</td>
<td>$-0.007$</td>
<td>$-0.007$</td>
<td>$-0.010$</td>
<td>$-0.010$</td>
</tr>
<tr>
<td>$\sqrt{MSE} \hat{d}$</td>
<td>$0.047$</td>
<td>$0.042$</td>
<td>$0.042$</td>
<td>$0.045$</td>
<td>$0.049$</td>
<td>$0.043$</td>
<td>$0.044$</td>
<td>$0.049$</td>
<td>$0.040$</td>
</tr>
<tr>
<td>$k = T^{1/4}$</td>
<td>$0.046$</td>
<td>$0.042$</td>
<td>$0.042$</td>
<td>$0.043$</td>
<td>$0.046$</td>
<td>$0.044$</td>
<td>$0.044$</td>
<td>$0.046$</td>
<td>$0.042$</td>
</tr>
<tr>
<td>$k = T^{1/2}$</td>
<td>$0.052$</td>
<td>$0.043$</td>
<td>$0.042$</td>
<td>$0.047$</td>
<td>$0.049$</td>
<td>$0.044$</td>
<td>$0.049$</td>
<td>$0.050$</td>
<td>$0.045$</td>
</tr>
</tbody>
</table>

* $T = 100$ * $T = 400$
For $T = 100$, the square root of the MSE is in general smaller for those estimates computed with $k = T^{1/4}$, although for $T = 400$ the latter quantity is very similar and even smaller for those computed with $T^{1/3}$. The standard deviations of $\hat{d}$ obtained by simulation are very good approximations to the asymptotic ones (obtained as $T^{-1/2}(\sum_{j=1}^{k} j^{-2})^{-1/2}$). They are also close to the asymptotically efficient standard deviations, computed as $T^{-1/2}(\sum_{j=1}^{\infty} j^{-2})^{-1/2}$, see (18), which equal 0.078 and 0.039 for $T = 100$ and 400, respectively. It is also remarkable that the estimates present a small bias, even for very large or very small values of $d$. The bias is usually negative which suggests that the GMD method slightly underestimates the memory parameter. Also note that the figures for the case where $\mu_0$ is unknown do not differ significantly from the case where it is known, as the theory predicts.

Tables 3–5 display the results obtained from similar experiments for the case where $u_t$ is an AR(1), MA(1) or an AR(2) process with parameters $\phi_1 = 0.6$, $\theta_1 = 0.5$ and $\phi_1 = 0.65$ and $\phi_2 = -0.6$, respectively. The values of the parameters have been chosen in accordance to previous studies to facilitate the comparison. Different values of $k$ were used and similar conclusions as in Tables 1 and 2 could be drawn. Hence, only the figures computed with $k = T^{1/4}$ are displayed. Also, since the results for the case where $\mu_0$ is known or unknown are very similar, only the latter are reported.

Finally, the GMD estimator has been compared to other competing methods. More specifically, we consider the Whittle estimator (with Zhurbenko taper of order 2) proposed by Velasco and Robinson (VR), the ML estimators introduced by Beran (BER) and Sowell (SOW) and the minimum distance estimator by Tieslau et al. (1996), (TSB). Table 6 reports the bias and the square root of the MSE for the above-mentioned estimators (the figures corresponding to the GMD are also reported to facilitate the comparison) for the case where the DGP is ARFIMA (0, $d$, 0) with known mean (equal to zero) and Gaussian innovations. The sample size is $T = 100$. The missing values in Table 6 stem from methods that are not defined for the whole range of values of $d$.

It can be seen that the VR (with tapering) and the TBS method are clearly inefficient with respect to the remaining methods. The ML estimators (BER and SOW) behave very similarly and slightly outperform in terms of MSE the GMD method for moderate values of $d$. For high values of $d$, both bias and MSE are smaller for the GMD estimator.

The relative advantage of the ML methods rapidly disappears when larger sample sizes or more general DGPs are considered. For instance, for $T = 400$ the $\sqrt{\text{MSE}}$ for the BER and SOW estimators is virtually the same as that of the GMD estimator reported in Table 1 (around 0.041). Also, when the mean is considered as unknown, the $\sqrt{\text{MSE}}$ of the ML estimators increases considerably. For the BER estimator, it is equal to 0.104, 0.102 for $d = \{0.2, 0.4\}$ and $T = 100$ while for the SOW case the corresponding figures are 0.112 and 0.105 while are higher in both cases as those of the GMD, reported in Table 2 (0.099 and 0.097, respectively).

Finally, non-Gaussian innovations have also been considered (more specifically, $\chi^2$ and $t$ distributions with 1 and 3 degrees of freedom). In this case, the GMD estimator clearly outperforms the former especially for moderate sample sizes.

---

8More specifically, for $T = 100$ asymptotic standard deviations are equal to 0.0857, 0.0827 and 0.8030 for $k = 3$, 5 and 10, respectively. For $T = 400$, the corresponding figures are 0.0419, 0.0407 and 0.396 for $k = 4$, 7 and 20, respectively.

9More complex ARMA models (containing up to 3 AR and MA lags) have also been considered and the results, not reported to save space, are available upon request. The estimator was also well-behaved in these more general cases although, not surprisingly, the variance increases as the number of parameters to be estimated gets larger.
Table 3. Estimation of $\lambda = (d, \phi_1)'$. DGP: ARFIMA (1, $d$, 0), $\phi_1 = 0.6$, $k = T^{1/4}$.

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>-0.7</th>
<th>-0.2</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
<th>1.4</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>-0.022</td>
<td>-0.021</td>
<td>-0.009</td>
<td>-0.008</td>
<td>0.000</td>
<td>-0.011</td>
<td>-0.045</td>
<td>-0.021</td>
<td>-0.023</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{d}$</td>
<td>0.219</td>
<td>0.221</td>
<td>0.171</td>
<td>0.164</td>
<td>0.181</td>
<td>0.219</td>
<td>0.221</td>
<td>0.221</td>
<td>0.233</td>
</tr>
<tr>
<td>Bias $\hat{\phi}_1$</td>
<td>-0.024</td>
<td>-0.043</td>
<td>-0.024</td>
<td>-0.031</td>
<td>-0.041</td>
<td>-0.025</td>
<td>0.003</td>
<td>0.000</td>
<td>0.010</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{\phi}_1$</td>
<td>0.211</td>
<td>0.197</td>
<td>0.201</td>
<td>0.0185</td>
<td>0.182</td>
<td>0.182</td>
<td>0.216</td>
<td>0.190</td>
<td>0.215</td>
</tr>
<tr>
<td>\text{bias $\hat{d}$}</td>
<td>-0.004</td>
<td>-0.031</td>
<td>-0.021</td>
<td>-0.017</td>
<td>-0.018</td>
<td>-0.024</td>
<td>-0.023</td>
<td>-0.013</td>
<td>-0.010</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{d}$</td>
<td>0.146</td>
<td>0.141</td>
<td>0.134</td>
<td>0.141</td>
<td>0.131</td>
<td>0.143</td>
<td>0.140</td>
<td>0.143</td>
<td>0.144</td>
</tr>
<tr>
<td>Bias $\hat{\phi}_1$</td>
<td>0.018</td>
<td>0.004</td>
<td>0.008</td>
<td>0.010</td>
<td>0.013</td>
<td>0.001</td>
<td>0.003</td>
<td>0.026</td>
<td>-0.016</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}} \hat{\phi}_1$</td>
<td>0.143</td>
<td>0.136</td>
<td>0.130</td>
<td>0.114</td>
<td>0.127</td>
<td>0.147</td>
<td>0.132</td>
<td>0.139</td>
<td>0.140</td>
</tr>
</tbody>
</table>
Bias $\hat{\lambda}$ introduced a goodness-of-fit procedure that tests for significant residual autocorrelations. For this reason, a formal test to check model adequacy is often carried out. A common way of the good properties of parametric estimators rely heavily on the correct specification of the model.

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>−0.7</th>
<th>−0.2</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
<th>1.4</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\lambda}$</td>
<td>−0.019</td>
<td>−0.021</td>
<td>−0.009</td>
<td>−0.010</td>
<td>−0.008</td>
<td>−0.016</td>
<td>−0.021</td>
<td>−0.018</td>
<td>−0.022</td>
</tr>
<tr>
<td>$\sqrt{MSE} \hat{\lambda}$</td>
<td>0.135</td>
<td>0.129</td>
<td>0.124</td>
<td>0.132</td>
<td>0.147</td>
<td>0.145</td>
<td>0.140</td>
<td>0.148</td>
<td>0.146</td>
</tr>
<tr>
<td>$\hat{\theta}_1$</td>
<td>0.002</td>
<td>0.007</td>
<td>0.008</td>
<td>−0.001</td>
<td>−0.005</td>
<td>−0.000</td>
<td>0.012</td>
<td>0.002</td>
<td>0.007</td>
</tr>
<tr>
<td>$\sqrt{MSE} \hat{\theta}_1$</td>
<td>0.129</td>
<td>0.125</td>
<td>0.124</td>
<td>0.129</td>
<td>0.135</td>
<td>0.136</td>
<td>0.134</td>
<td>0.135</td>
<td>0.139</td>
</tr>
</tbody>
</table>

Table 5. Estimation of $\lambda = (d, \phi_1, \phi_2)'$. DGP: ARFIMA(2, d, 0), $\phi_1 = 0.6$, $\phi_2 = −0.65$; $k = T^{1/4}$.

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>−0.7</th>
<th>−0.2</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
<th>1.4</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>bias $\hat{\lambda}$</td>
<td>−0.003</td>
<td>−0.031</td>
<td>−0.028</td>
<td>−0.029</td>
<td>−0.030</td>
<td>−0.021</td>
<td>−0.022</td>
<td>−0.064</td>
<td>−0.168</td>
</tr>
<tr>
<td>$\sqrt{MSE} \hat{\lambda}$</td>
<td>0.224</td>
<td>0.228</td>
<td>0.211</td>
<td>0.223</td>
<td>0.230</td>
<td>0.229</td>
<td>0.223</td>
<td>0.225</td>
<td>0.281</td>
</tr>
<tr>
<td>bias $\phi_1$</td>
<td>0.027</td>
<td>0.074</td>
<td>0.029</td>
<td>0.031</td>
<td>0.052</td>
<td>0.031</td>
<td>0.041</td>
<td>0.058</td>
<td>0.049</td>
</tr>
<tr>
<td>$\sqrt{MSE} \phi_1$</td>
<td>0.206</td>
<td>0.219</td>
<td>0.201</td>
<td>0.221</td>
<td>0.232</td>
<td>0.216</td>
<td>0.226</td>
<td>0.230</td>
<td>0.302</td>
</tr>
<tr>
<td>bias $\phi_2$</td>
<td>−0.059</td>
<td>−0.050</td>
<td>−0.058</td>
<td>−0.049</td>
<td>−0.083</td>
<td>−0.069</td>
<td>−0.072</td>
<td>−0.067</td>
<td>−0.031</td>
</tr>
<tr>
<td>$\sqrt{MSE} \phi_2$</td>
<td>0.150</td>
<td>0.171</td>
<td>0.151</td>
<td>0.171</td>
<td>0.174</td>
<td>0.159</td>
<td>0.165</td>
<td>0.165</td>
<td>0.371</td>
</tr>
</tbody>
</table>

5. RESIDUAL-BASED STATISTICS FOR DIAGNOSING CHECKING

The good properties of parametric estimators rely heavily on the correct specification of the model. For this reason, a formal test to check model adequacy is often carried out. A common way of testing this hypothesis is by checking the assumption of white noise residuals. Box and Pierce (1970) introduced a goodness-of-fit procedure that tests for significant residual autocorrelations. The Box–Pierce (BP) statistic is defined as

$$Q(k) = T \sum_{i=1}^{k} \hat{\phi}_1^2(i).$$

© Royal Economic Society 2007
Table 6. Bias and $\sqrt{\text{MSE}}$ for various estimation techniques DGP: $\Delta^{d_0} y_t = \varepsilon_t$; $\mu_0$ known.

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>Bias</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
<th>1.4</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>VR</td>
<td>-0.174</td>
<td>-0.155</td>
<td>-0.108</td>
<td>-0.129</td>
<td>-0.090</td>
<td>-0.064</td>
<td>0.027</td>
<td>0.100</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}}$</td>
<td>0.232</td>
<td>0.218</td>
<td>0.201</td>
<td>0.193</td>
<td>0.172</td>
<td>0.155</td>
<td>0.154</td>
<td>0.160</td>
</tr>
<tr>
<td>BER</td>
<td>-0.005</td>
<td>-0.010</td>
<td>-0.012</td>
<td>0.020</td>
<td>-0.005</td>
<td>-0.013</td>
<td>-0.037</td>
<td>-0.157</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}}$</td>
<td>0.083</td>
<td>0.085</td>
<td>0.084</td>
<td>0.087</td>
<td>0.083</td>
<td>0.08</td>
<td>0.093</td>
<td>0.212</td>
</tr>
<tr>
<td>TSB</td>
<td>-0.121</td>
<td>-0.041</td>
<td>-0.094</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}}$</td>
<td>0.189</td>
<td>0.094</td>
<td>0.109</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SOW</td>
<td>-0.011</td>
<td>-0.011</td>
<td>-0.012</td>
<td>-0.020</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}}$</td>
<td>0.089</td>
<td>0.084</td>
<td>0.080</td>
<td>0.084</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GMD</td>
<td>-0.025</td>
<td>-0.020</td>
<td>-0.018</td>
<td>-0.021</td>
<td>-0.017</td>
<td>-0.024</td>
<td>-0.019</td>
<td>-0.014</td>
</tr>
<tr>
<td>$\sqrt{\text{MSE}}$</td>
<td>0.101</td>
<td>0.092</td>
<td>0.090</td>
<td>0.094</td>
<td>0.096</td>
<td>0.097</td>
<td>0.093</td>
<td>0.104</td>
</tr>
</tbody>
</table>

They showed that, in the context of ARIMA processes, this statistic is asymptotically $\chi^2$-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 $\chi^2_{k-p-q}$ distribution. This statistic is defined as

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

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

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

where $h(.)$ is a suitable 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 that

$$H^*_T = \left( T \sum_{i=1}^{kT} \hat{p}^2_{\varepsilon(i)}(i) - k(T) \right) / (2k(T))^{1/2},$$

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 behaviour 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 extends Li and McLeod’s
result to the case where the process is allowed to be non-stationary and to have an unknown mean. The same distribution is found to hold in this more general case.

**Theorem 4** Let $\hat{\rho}_{e(\lambda)}$ be the vector containing the autocorrelations up to lag $k$ of the residuals, defined in (5) or (7), such that

$$
\hat{\rho}_{e(\lambda)} = \left( \hat{\rho}_{e(\lambda)}(1), \ldots, \hat{\rho}_{e(\lambda)}(k) \right)^T,
$$

where $\hat{\rho}_{e(\lambda)}(i)$ is defined as in (10). Then, for any fixed $k$, $\sqrt{T} \hat{\rho}_{e(\lambda)}$ 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 $\Sigma$ is approximately idempotent with rank $k - p - q - 1$. Hence, both $Q(k)$ and $\tilde{Q}(k)$ converge to a $\chi^2$ distribution with $k - p - q - 1$ degrees of freedom, for $k$ strictly greater than $p + q + 1$, as $T$ goes to infinity.\(^\text{10}\)

The previous results imply that the minimum value of the criterion function (13), $V_{ke}(\hat{\lambda})$, 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 as a $\chi^2_{k-p-q-1}$ distribution for large $T$, where $k$ is the number of autocorrelations considered in the criterion function.\(^\text{11}\)

The following section explores by simulation the finite sample properties of the statistics $Q(k)$ and $\tilde{Q}(k)$.

### 6. BEHAVIOUR 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 $\Delta^{d_0} y_t = \epsilon_t$, $\epsilon_t \sim NID(0, 1)$ 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% significance level is calculated using the $\chi^2_{k-p-q-1,0.95}$ value. Since the Ljung–Box statistic improves the approximation to the $\chi^2$ 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 behaviour of both tests. Table 7 reports the empirical size of both the Box–Pierce (BP) and Ljung–Box (LB) tests. In agreement with

\(^{10}\)Note that for the ARFIMA(0,d,0) model if $k = 1$, the distribution described in Theorem 4 has zero variance and then $\hat{\rho}_{e(\lambda)}(1) \xrightarrow{p} 0$. Simulation experiments corroborate this result (for instance, for $T = 400$, the mean of $\hat{\rho}_{e(\lambda)}(1)$ is 0.0045 and the variance is of an order of magnitude of $10^{-5}$).

\(^{11}\)It would also be straightforward to compute the Hong’s statistic in (20) from $V_{ke}(\hat{\lambda})$ just by multiplying it by the sample size, subtracting the number of included autocorrelations and dividing by the squared root of $2k(T)$. Nevertheless, since the asymptotic properties of this statistic are unknown when long memory processes are considered, we do not explore its properties any further.
Minimum distance estimation

Table 7. Empirical Size of BP and LB (S.L.:5%) DGP: ARFIMA(0, d, 0).
\[d_0\] -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%

the findings of Li and McLeod (1986), the empirical size is close to the nominal size in both tests. Although the approximation to the \(\chi^2_{k-1}\) 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) has been chosen as the true DGP with a value of the autoregressive parameter \(\phi_1\) equal to 0.5, for different values of \(d_0\). An ARFIMA (0, \(d_0\), 0) was estimated instead. Table 8 reports the power of both the BP and LB tests at the 5% nominal signification level.\(^{12}\) Both tests perform quite similarly. Also, for small sample sizes (\(T = 150\)) power is similar to that obtained by other methods proposed in the literature (see, e.g. Delgado and Hidalgo 1999). The rejection rates improve considerably when larger sample sizes are considered, providing good results for sample sizes around 400 or 500.

\(^{12}\)Note the figures representing power are not size-adjusted. Since both tests are in general undersized, higher power would have been achieved after size-adjustment.
7. EMPIRICAL APPLICATIONS

To illustrate the application of the technique proposed in this paper and to provide further comparison with previous estimation techniques, the series analyzed by Beran (1995), Velasco and Robinson (2000) and Robinson (1994) have been estimated, namely the chemical process temperature readings (Series C) and the chemical process concentration readings (Series A) from Box and Jenkins (1976).

Our conclusions are in fair agreement with those obtained in 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 to the GMD method are $\hat{d} = 1.005$ with a 95% confidence interval (CI) of [0.7497, 1.2617] and $\hat{\phi}_1$ is 0.798 with a 95% CI of [0.563, 0.972], in close agreement with BJ’s conclusions. Nevertheless, the recognition of the uncertainty on $d$ substantially increases the standard deviation of the AR parameter. Similar conclusions are reached both in Beran (1995) and in Velasco and Robinson (2000).

For series A, BJ fitted an ARIMA(0, 1, 1), which yields a significant value for the MA parameter equal to $-0.7$. This large negative value suggests that the model could be over differenced. If an ARFIMA(0, $d$, 1) is fitted instead, the GMD estimates are 0.43 and $-0.038$ with 95% CI 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 CI, which reinforces the belief 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% CI of [0.292, 0.51]. Also in this case, our results follow closely those obtained in Velasco and Robinson (2000) and Beran (1995).

8. CONCLUSIONS

In this paper, we have proposed a new method for estimating the parameters of an ARFIMA($p$, $d_0$, $q$) process with $d_0 > -0.75$. 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 residuals 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. 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 estimated values coincides with the Box and Pierce (1970) goodness-of-fit statistic, providing, therefore, an 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 heteroscedasticity 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
Minimum distance estimation

stochastic volatility model. Further research should also be undertaken to extend the previous framework to the multivariate case.

ACKNOWLEDGEMENTS

I am very grateful for considerable help and suggestions from Juan J. Dolado. I would also like to thank the editor, two anonymous referees, Gabor Lugosi, Francesc Mármol, Peter Phillips and Carlos Velasco for their valuable comments. I also acknowledge financial support from the Spanish Ministry of Science and Technology through grants SEC2003-04429 and SEC2003-04476 and the Barcelona Economics Program of CREA. All remaining errors are, of course, my own.

REFERENCES


© Royal Economic Society 2007

**APPENDIX**

**Proof of Theorem 1.** The residuals $e_t(\lambda)$ in (7) evaluated at $\lambda = \lambda_0$ can be written as

\[ e_t(\lambda_0) = \sum_{j=0}^{t-m_0-1} \alpha_j \left( \Delta^{m_0} y_{t-j} - \mu_0 \right) + \left( \Delta^{m_0} \bar{y} - \mu_0 \right) \sum_{j=0}^{t-m_0-1} \alpha_j \left( \lambda_0^* \right), \quad t = m_0 + 1, \ldots, T. \]

This allows us to write

\[ e_t(\lambda_0) = \varepsilon_t + \eta_t, \]
where $\eta_t$ is given by

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

and $\Delta^{m_0} \bar{y}$ 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,

$$\sqrt{T} \Upsilon_T = \sqrt{T} (\hat{\rho}_{x(\epsilon_0)} - \hat{\rho}_{x(i)}) ,$$

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

$$T^{-1/2} \left( \sum_{t=1}^{T-i} \varepsilon_{t+i} \eta_t + \sum_{t=1}^{T-i} \varepsilon_i \eta_{t+i} + \sum_{t=1}^{T-i} \eta_t \eta_{t+i} \right). \quad (A.1)$$

Let us first consider the case where $\mu_0$ is known and equal to zero. In this case, $\eta_t$ collapses to $\eta_t = -\sum_{i=0}^{\infty} \alpha_i (\lambda_0^*) y_{t-i}$. By repeated substitution, Odaki (1993) shows that this expression can be rewritten as

$$\eta_t = \sum_{j=0}^{\infty} \psi_{j,t-1} \varepsilon_{-j} , \quad (A.2)$$

[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

$$\left( \sum_{j=0}^{\infty} \psi_{j,t-1}^2 \right) = \begin{cases} O (t^{-1}) \text{ if } \varphi_0 \in (-0.5, 0.5), \\ O ((\log t) t^{-1}) \text{ if } \varphi_0 = -0.5, \\ O (t^{-(2(1+\varphi_0))}) \text{ if } \varphi_0 < -0.5. \end{cases} \quad (A.3)$$

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

$$T^{-1} E \left( \sum_{t=1}^{T-i} \varepsilon_{t+i}^2 \eta_t^2 \right) = \sigma^4 T^{-1} \sum_{t=1}^{T-i} \sum_{j=0}^{\infty} \psi_{j,t-1}^2 \rightarrow 0, \quad (A.4)$$

since it follows from (A.3) that

$$\sum_{t=1}^{T-i} \left( \sum_{j=0}^{\infty} \psi_{j,t-1}^2 \right) = \begin{cases} O ((\log(T - i))) \text{ for } \varphi_0 \in (-0.5, 0.5) \\ O ((\log(T - i))^2) \text{ for } \varphi_0 = -0.5 \\ O ((T - i)^{-2(1+\varphi_0)+1}) \text{ for } \varphi_0 < -0.5. \end{cases} \quad (A.5)$$

Since $\sum_{j=0}^{\infty} E(\psi_{j,t-1}^2 \varepsilon_{-j}^2) < \infty$, it is possible to interchange the expectation and the summation in order to obtain (A.4) (See Rao (1973) p. 111).
A similar argument holds for the second term in (A.1), since in this case \( \eta_{i+i} = \sum_{j=0}^{\infty} \psi_{j,j+i-1} \varepsilon_j \) and again no contemporaneous terms of \( \varepsilon \) are found in the product \( \varepsilon_i \eta_{i+i} \). With respect to the third term in (A.1),

\[
T^{-1/2} \sum_{i=1}^{T-1} E \left| \eta_i \eta_{i+i} \right| \leq T^{-1/2} \sum_{i=1}^{T-1} E \left( \eta_i^2 \right) = T^{-1/2} \sigma^2 \sum_{i=1}^{T-1} \left( \sum_{j=0}^{\infty} \psi_{j,j-1} \right)^p \to 0. \tag{A.6}
\]

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

\[
T^{-1} E \left( \sum_{i=1}^{T-i} \eta_i \eta_{i+i} \right)^2 \leq 2T^{-1} \sum_{r=1}^{T-i} E \left( \eta_i \eta_{r+i} \eta_s \eta_{r+i} \right), \tag{A.7}
\]

\[
E \left( \eta_i \eta_{r+i} \eta_s \eta_{r+i} \right) \leq \mu_4 \sum_{j=0}^{\infty} |\psi_{j,r-1} \psi_{j,r+i-1} \psi_{j,s-1} \psi_{j,s+i-1}| + \sigma^4 \sum_{j=0}^{\infty} |\psi_{j,r-1} |^2 \sum_{j=0}^{\infty} |\psi_{j,s-1} |^2 + 3\sigma^4 \sum_{j=0}^{\infty} \psi_{j,r-1}^2 \sum_{j=0}^{\infty} \psi_{j,s-1}^2. \tag{A.8}
\]

It can be checked that the coefficients \( \{ \psi_{j,t} \} \) are strictly decreasing in both subindexes \( (j, t) \). Using these results and Cauchy’s inequality, it is obtained that for \( s \geq r \),

\[
E \left( \eta_i \eta_{r+i} \eta_s \eta_{r+i} \right) \leq \mu_4 \left( \sum_{j=0}^{\infty} \psi_{j,r-1}^4 \right)^{1/2} \left( \sum_{j=0}^{\infty} \psi_{j,s-1}^4 \right)^{1/2} + 3\sigma^4 \sum_{j=0}^{\infty} \psi_{j,r-1}^2 \sum_{j=0}^{\infty} \psi_{j,s-1}^2. \tag{A.9}
\]

Following Odaki (1993, p. 707) it is easy to show that \( \sum_{j=0}^{\infty} \psi_{j,t}^4 = O(t^{-3}) \) for \( d > -0.75 \). Taking this result into account and the orders of magnitude in (A.3), it follows that

\[
T^{-1} \sum_{r=1}^{T-i} \sum_{s \geq r} E \left( \eta_i \eta_{r+i} \eta_s \eta_{r+i} \right) \leq T^{-1} \left[ \mu_4 \sum_{r=1}^{T-i} \left( \sum_{j=0}^{\infty} \psi_{j,r-1}^4 \right)^{1/2} \sum_{s=1}^{T-i} \left( \sum_{j=0}^{\infty} \psi_{j,s-1}^4 \right)^{1/2} + 3\sigma^4 \sum_{r=1}^{T-i} \left( \sum_{j=0}^{\infty} \psi_{j,r-1}^2 \right) \sum_{s=1}^{T-i} \left( \sum_{j=0}^{\infty} \psi_{j,s-1}^2 \right) \right] \to 0.
\]

for \( \varphi_0 > -0.75 \). Since the \( i \)th element of \( \sqrt{T} \gamma_T \) tends to zero for all \( i = 1, \ldots, k \), then \( \sqrt{T} \gamma_T \xrightarrow{p} 0 \), implying the desired result.

The case where \( \mu_0 \) is unknown can be proved similarly using standard arguments since \( \Delta^{m,1} y \) is a consistent estimator of \( \mu_0 \) (see Robinson, 1994).

\( \square \)

In order to prove the consistency of the estimator stated in Theorem 2, we consider separately the cases where the inferior limit of the parameter space of \( d \), \( \nu_1 \), is such that \( \nu_1 \leq d_0 - 1/2 \) from those where \( \nu_1 > d_0 - 1/2 \), where \( d_0 \) is the true integration order of the process to be estimated, \( Y_t \). The reason for making this distinction is the non-uniform behaviour of FI(\( \delta \)) processes which determines the properties of the criterion function. More specifically, whenever \( \delta \geq 1/2 \), correlations are not well-defined since the process is not stationary. To define the residuals, \( \varepsilon_t(d) \), the process \( y_t \) is filtered by \( \Delta^d \), so that \( \varepsilon_t(d) \) is a FI(\( d_0 - d \)) process. If \( \nu_1 > d_0 - 1/2 \), then \( d_0 - d < 1/2 \) for all \( d \in [\nu_1, \nu_2] \), and the autocorrelations of
the process \( e_t(d) \) 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 similar problem appeared in Robinson (1995) and Velasco and Robinson (2000). To overcome it, they proposed a two-step proof that will also be 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 a subset where the correlations are well-defined is consistent, while the second proves that values in the remaining set 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 A.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 A.1** Let \( V_{k_e}(\lambda) \) 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 \( \lambda \in \Lambda_1 \) and define \( V_{k_e}(\lambda) = \sum_{k=1}^{K}(p_{\lambda}(i))^2 \), where \( p_{\lambda}(i) \) are the (population) autocorrelations associated to the non-truncated residuals \( \epsilon(\lambda) = \sum_{j=0}^{\infty} \rho_{\lambda}(j) \epsilon(m_{\lambda,j}) \), and \( k \) is a fixed number. Then:

1. \( V_{k_e}(\lambda) \) is continuous in \( \lambda \), \( V_{k_e}(\lambda) \) converges in probability to \( V_{k_e}(\lambda) \) and the convergence is uniform.
2. \( V_{k_e}(\lambda) \) is a continuous function and has a unique minimum at \( \lambda_0 \), such that \( V_{k_e}(\lambda_0) = 0 \).

**Proof of Lemma 1.**

1. The continuity of \( V_{k_e}(\lambda) \) 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 \( V_{k_e}(\lambda) \) converges in probability and an equicontinuity argument using the compactness of \( \Lambda_1 \) and the differentiability of \( \rho(\cdot \lambda) \) with respect to \( \lambda \) (cf. Davidson 1994, p. 340, and Velasco and Robinson 2000).
2. It is straightforward to check that \( V_{k_e}(\lambda) \) has a unique minimum at \( \lambda_0 \), since \( e_t(\lambda)|_{\lambda=\lambda_0} = e_t \), which is an i.i.d. process and therefore all its correlations are zero (which implies that \( V_{k_e}(\lambda_0) = 0 \), but presents non-null autocorrelations for any other value of \( \lambda \neq \lambda_0 \), and therefore \( V_{k_e}(\lambda) > 0 \). The continuity of \( V_{k_e}(\lambda) \) follows from the assumptions above (see Amemiya 1985, theorem 4.1.1).

**Proof of Theorem 2.** We proceed with the two-step proof of consistency proposed in Robinson (1995) and Velasco and Robinson (2000).

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

\[
\sup_{\lambda \in \Lambda_1} \left| U(\lambda) \right| \leq \sup_{\lambda \in \Lambda_1} \left| V_{k_e}(\lambda) - V_{k_e}(\lambda) \right| + \left| V_{k_e}(\lambda_0) \right|, \tag{A.10}
\]

and both terms in the right-hand side of (A.10) tend to zero, the first due to uniform convergence and the second due to pointwise convergence (Lemma 1).

**Second step.** Recall that \( \Lambda_1 = \{d : \nabla_1 \leq d \leq \nabla_2 \} \times \Lambda^{(-1)} \), 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 \( \Lambda_2 = \{d : \nabla_1 < d < d_0 - 1/2 + \eta \} \times \Lambda^{(-1)} \) and let \( \hat{\lambda}_k = \arg \min_{\lambda \in \Lambda} V_{k_e}(\lambda) \) be the estimator.
computed in the whole parametric space, \( \Lambda = \Lambda_1 \cup \Lambda_2 \). We want to show that \( \hat{\lambda}_k - \hat{\lambda}_1 \overset{p}{\to} 0 \) or, equivalently, that for any \( \delta > 0 \), \( P(\|\hat{\lambda}_k - \hat{\lambda}_1\| \geq \delta) \to 0 \). Note that,

\[
P \left( \|\hat{\lambda}_k - \hat{\lambda}_1\| \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( \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq V_{ke}(\hat{\lambda}_1) \right) \cap \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0 \) \quad \text{(A.11)}
\]

\[
+ P \left( \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq V_{ke}(\hat{\lambda}_1) \right) \cap \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) > 0 \)
\]

\[
\leq P \left( \inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0 \right) + P \left( V_{ke}(\hat{\lambda}_1) > 0 \right). \quad \text{(A.12)}
\]

Since \( \hat{\lambda}_1 \) is consistent (step 1) and uniform convergence of \( V_{ke} \) holds, the second probability in (A.12) 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 \( \Lambda_2 \), contains the squared sample correlations of a FI\((d_0 - \lambda^{(1)})\) process, where \( \lambda^{(1)} \in [\lambda_1, d_0 - 1/2 + \eta] \), 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 correlations converge to the squared population autocorrelations which, clearly, are bounded away from zero. If \( \lambda^{(1)} \in [\lambda_1, d_0 - 1/2] \), 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 \( V_{ke}(\lambda) = \sum_{i=1}^{k} \hat{\rho}^2_{\text{est}}(i) \geq \hat{\rho}^2_{\text{est}}(1) \overset{p}{\to} 1 \) (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 \( P(\inf_{\lambda \in \Lambda_2} V_{ke}(\lambda) \leq 0) \to 0 \).

It follows that the estimator \( \hat{\lambda}_k \) computed in the whole parametric space, \( \Lambda \), has the same limit as \( \hat{\lambda}_1 \), calculated in the restricted parametric space \( \Lambda_1 \). Since the latter estimator was shown to be consistent in the first part of the proof, it follows that \( \hat{\lambda}_k \) is also a consistent estimator of \( \lambda_0 \).

\[
\square
\]

In order to derive the asymptotic distribution of \( \hat{\lambda}_k \), the following auxiliary result is required.

**Lemma A.2.** Let \( J_\lambda(\lambda) = \partial_\lambda (\hat{\rho}_{\text{est}}(\lambda)^T) \) be the Jacobian matrix of \( \hat{\rho}_{\text{est}}(\lambda) \), that is, the \( k \times (p + q + 1) \) matrix of partial derivatives of \( \hat{\rho}_{\text{est}}(\lambda) \) with respect to \( \lambda \). Under the hypotheses of Sections 2 and 3, it holds that

1. for each finite \( k \),

\[
\hat{J}_k(\lambda_0) \overset{p}{\to} J_k(\lambda_0),
\]

where

\[
J_k(\lambda_0) = \begin{pmatrix}
-1 & 1 & 0 & \ldots & 1 & \ldots & 0 \\
-1/2 & \omega_1 & 1 & \ldots & \psi_1 & \ldots & 0 \\
& \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
-1/k & \omega_{k-1} & \omega_{k-2} & \ldots & \psi_{k-1} & \ldots & \psi_{k-q}
\end{pmatrix} \quad \text{(A.13)}
\]

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

2. The first derivative of \( \hat{J}_\lambda(\lambda) \) exists and is bounded in probability in an open convex set containing \( \lambda_0 \).

**Proof of Lemma 2.**

1. This result can be found in theorem 2 in Li and McLeod (1986).
2. Let \( \tilde{A} \subset \Lambda_1 \) be an open convex set containing \( \lambda_0 \). Note that \( \hat{J}_\lambda(\lambda) \) is differentiable in this set since it is a function of the derivatives of the residuals \( e_r(\lambda) \) which are \( C(\infty) \) with respect to \( \lambda \) in the relevant set. The first derivative of \( \hat{J}_\lambda(\lambda) \) is the \( k(p + q + 1) \times (p + q + 1) \) matrix defined as

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

$$\hat{\rho}''_{e(1,1)}(1) = \frac{\sigma_{e(1,1)}^2}{\sum e_i^2(\lambda)/T} \left\{ \hat{\rho}''_{e(1,1)}(1) - \hat{\rho}'_{e(1,1)}(1) \left[ \frac{4 \sum e_i e_i' (\lambda)}{\sum e_i^2(\lambda)} \right] \right\}, \quad (A.14)$$

where

$$\hat{\rho}_{e(1,1)}(1) = \sigma^{-2} \sum_{t=1}^{T-j} e_i(\lambda)e_{t+1}(\lambda)/T, \quad \sigma_{e(1,1)}^2 = \lim_{T \to \infty} T^{-1} \sum e_i^2(\lambda),$$

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

and

$$\hat{\rho}''_{e(1,1)}(1) = \frac{\sum e_i'' (\lambda) e_{t+1}(\lambda) + \sum e_i' (\lambda) e_{t+1}'(\lambda) + 2 \sum e_i' (\lambda) e_{t+1}'(\lambda)}{T}. \quad (A.16)$$

Noting that $e_i = (1 - (L)^j \Theta^{-1} \Phi(L)_{1,0.2}^j$ for large $j$, one can construct along the same lines and for the sake of brevity are also omitted.

Each of the terms in the RHS of (A.14) and (A.15) is $O_p(1)$. Consider for instance $\hat{\rho}'_{e(1,1)}(1)$. The first term in (A.16) can be rewritten as

$$\frac{T}{\sum (\log^2 (1 - L)e_{t+1}(d))e_i(d)} = \frac{T}{\sum [\zeta_2 e_{i-1}(d) + \zeta_3 e_{i-2}(d) + \zeta_4 e_{i-3}(d) + \cdots + \zeta_r e_{i}(d)] e_i(d)}$$

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

$$\zeta_{j+11}(j) \approx C \frac{2}{j + 1} \left( \sum_{i=1}^{j} i^{-1} \right) j^{2(d_0-d)-1} \approx C' j^{2(d_0-d)-2} \log j$$

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 $\lim_{T \to \infty} \sum_{j=1}^{T} \zeta_{j+1}(j)$ is bounded. On the other hand, since $\hat{\gamma}(j)$ is a consistent estimator of the covariance function, then

$$\lim_{T \to \infty} \sum_{j=1}^{T} \zeta_{j+1}(j) = O_p(1).$$

The proof for the remaining components in (A.14) and (A.15) is analogous and therefore is omitted. Similarly, proofs for the remaining elements of the matrix $\partial vec(\partial \hat{\rho}_{e(1,1)}/\partial \lambda)$ can be constructed along the same lines and for the sake of brevity are also omitted.
**Proof of Theorem 3.** The mean value theorem applied to the first-order condition gives
\[ 0 = \frac{\partial V_{k\varepsilon}(\lambda)}{\partial \lambda} + \frac{\partial^2 V_{k\varepsilon}(\tilde{\lambda})}{\partial \lambda \partial \lambda'} (\tilde{\lambda}_k - \lambda_0), \]
where \( \tilde{\lambda} \) belongs to the line joining \( \tilde{\lambda}_k \) and \( \lambda_0 \). Multiplying through by \( \sqrt{T} \) and solving for \( \sqrt{T}(\tilde{\lambda}_k - \lambda_0) \) yield
\[ \sqrt{T}(\tilde{\lambda}_k - \lambda_0) = -\left( \frac{\partial \hat{\rho}'(\lambda_0)}{\partial \lambda} \frac{\partial \hat{\rho}'(\lambda_0)}{\partial \lambda'} \right)^{-1} \frac{\partial \hat{\rho}'(\lambda_0)}{\partial \lambda} \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda_0). \]
(A.17)
Further, \( \sqrt{T}(\tilde{\lambda}_k - \lambda_0) \) is asymptotically normally distributed, with variance–covariance matrix given by \( \Sigma_k^{-1} = (J_k(\lambda_0)'J_k(\lambda_0))^{-1} \).

**Proof of Theorem 4.** Let \( \hat{\rho}_{k\varepsilon}(\lambda) \) be the vector defined in (21). A first-order Taylor’s series expansion of \( \hat{\rho}_{k\varepsilon}(\lambda) \) around \( \lambda_0 \) yields
\[ \hat{\rho}_{k\varepsilon}(\lambda) = \hat{\rho}_{k\varepsilon}(\lambda_0) + \frac{\partial \hat{\rho}_{k\varepsilon}(\lambda_0)}{\partial \lambda} (\hat{\lambda} - \lambda_0). \]
(A.19)
By Theorem 1, \( \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda_0) = \sqrt{T} \hat{\rho}_{k\varepsilon} + o_p(1) \overset{w}{\to} N(0, I_k) \) for fixed \( k \) (Theorem 1), it is straightforward to check that \( \sqrt{T}(\tilde{\lambda}_k - \lambda_0) \) is also asymptotically normally distributed with mean equal to zero and variance–covariance matrix given by \( \Sigma_k^{-1} = (J_k(\lambda_0)'J_k(\lambda_0))^{-1} \).

Combining equations (A.19) and (A.17),
\[ \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda) = \sqrt{T} (I_k - Q) \hat{\rho}_{k\varepsilon}(\lambda_0) + o_p(1), \]
(A.20)
where \( Q = J_k(\lambda_0)'J_k(\lambda_0)^{-1} J_k(\lambda_0) \) is an idempotent matrix. It follows that
\[ \left( -\frac{Q}{I_k} \right) \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda_0) \]
is asymptotically normally distributed, with variance–covariance matrix given by
\[ \Sigma = \begin{pmatrix}
J_k(\lambda_0)'J_k(\lambda_0)^{-1} J_k(\lambda_0) & -J_k(\lambda_0)'J_k(\lambda_0)^{-1} J_k(\lambda_0) \\
-J_k(\lambda_0)'J_k(\lambda_0)^{-1} J_k(\lambda_0) & I_k
\end{pmatrix}. \]
Since joint normality holds, any linear combination of \( -\frac{Q}{\sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda_0)}, \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda_0) \) is also normally distributed. Hence, taking into account expression (A.20), it follows that
\[ \sqrt{T} \hat{\rho}_{k\varepsilon}(\lambda) \overset{w}{\to} N \left( 0, I_k - J_k(\lambda_0)'J_k(\lambda_0)^{-1} J_k(\lambda_0) \right). \]