Semi Parametric Estimation of Long Memory: Comparisons and Some Attractive Alternatives

Richard T. Baillie*
Departments of Economics and Finance, Michigan State University
Department of Economics, Queen Mary University of London

George Kapetanios
Department of Economics, Queen Mary University of London

Abstract

This paper focuses on the applicability of the semi parametric Local Whittle (LW) estimator of the long memory parameter in a univariate time series, when it is desired to estimate short run dynamics and the Impulse Response Weights (IRWs) of the time series model. The Local Whittle Two Step Estimator ($LWTSE$) of the long memory parameter and the subsequent estimates of short run parameters from fractionally filtered series are analyzed theoretically and by simulation. The $LWTSE$ is shown to generally have quite poor properties in the presence of moderately persistent autocorrelation in the short run component. These difficulties are also evident in the properties of the estimated IRWs. The alternative approach of using a time domain MLE to jointly estimate the parameters of the model and subsequently the IRWs is shown to be generally superior to using the $LWTSE$. An alternative purely non parametric approach is to use a high order linear autoregression (AR) approximation. This procedure is shown to provide consistent estimates of the infinite autoregressive approximation. A very attractive feature of the AR method is that it is valid for non stationary regions of the parameter space. In general, the high order AR method appears to provide relatively good estimates of IRW both for stationary and nonstationary processes and is nonparametric for both the long and short memory parts of the process.

Key Words: Long Memory, Local Whittle estimator, Non-stationarity, Maximum Likelihood Estimation, Autoregressive approximations.

JEL Codes: C22, C12.

*An earlier version of this paper was presented at the conference in honor of Peter Phillips' sixtieth birthday at Singapore Management University. The authors gratefully acknowledge very helpful comments from the conference participants at the Singapore Management University, the "International Symposium on Financial Econometrics", at University of Durham Business School, the EC2 conference "Advances in Time Series Econometrics", at the University of the Algarve, Portugal, and ESEM at Milano, Italy. Helpful comments were also received from Tim Bollerslev, Luidas Giraitis, Hira Koul, Peter Phillips, Peter Schmidt and Andrew Tremayne;. Any remaining errors are our own.

1 Introduction

Research on long memory and fractionally integrated processes has continued at an accelerating rate since the initial publication of the work of Granger and Joyeux (1980), Granger (1980) and Hosking (1981), which parameterized the processes of Hurst (1951) on time series with hyperbolically decaying autocorrelations. Some of the work on long memory models has been found to be extremely relevant for the theoretical modeling of volatility processes in financial market returns and for a variety of other asset pricing applications. However, and perhaps surprisingly, significantly more research effort appears to have been expended on the issue of semi parametric estimation (SPE) of the single long memory parameter in the conditional mean of a univariate time series. In fact, casual reading of recent literature would suggest that this is considered to be the most important problem in the long memory literature. The first SPE, known as the GPH estimator, was proposed by Geweke and Porter Hudak (1983), while the most widely used SPE now appears to be the Local Whittle estimator, which has been intensively investigated, e.g. see Robinson (1995) and Dalla, Giraitis, and Hidalgo (2005). There are also a plethora of new $SPEs$ of the long memory parameter, e.g. see Phillips (2007) and Phillips and Shimotsu (2006), which offer extensions and improvements of the $SPEs$ in terms of the range of the long memory parameter being covered, and in terms of bias reduction and choice of optimal bandwidth; e.g. Henry and Robinson (1996) and Henry (2001).

The point of departure of this paper is to question the paradigm of focusing virtually exclusive attention on the estimation of just the long memory parameter. The arguments presented here hinge on the fact that the single long memory parameter is merely one aspect of a model of a univariate time series and that the investigator is generally interested in wider issues. In particular, virtually no attention appears to have been given in the literature concerning how the SPE of the long memory parameter can be used in subsequent econometric work with the time series. In most practical situations the investigator will be interested in (i) building a time series model involving both the short and long memory components of the series, (ii) estimating Impulse Response Weights ($IRWs$) for a variety of lags, (iii) forming predictions over short and long horizons and (iv) using information from the univariate time series analysis for the construction of some multivariate econometric model.

The results presented in this paper indicate that SPE will generate very poor quality estimates of the long memory parameter when the short memory component has moderately persistent autocorrelation. The subsequent use of the SPE of the long memory parameter in this situation to generate second stage estimates of the short memory component and parameters, and then subsequently of the IRWs also turns out to have poor properties compared with a benchmark alternative method of MLE. The claim that the SPE approach may be preferred on the grounds that it does not require specification of the short run dynamics does not appear to be correct on the basis of the simulation results reported in this study. Nor does there seem to be any evidence that the MLE of $ARFIMA$ model parameters may be problematic due to potential high correlation between the estimates of the long memory and short memory parameters.

A further method considered in this paper which is completely nonparametric, as opposed to the SPE, is to approximate the long memory process by a truncated linear autoregression. The high order autoregressive (AR) approximation is shown to have very desirable theoretical properties and is shown to be valid over non stationary regions of the long memory parameter space.

The paper is organized as follows. Section 2 contains an outline of the basic model with long memory and short run dynamics. The arguments are presented in terms of the basic vanilla $ARFIMA$ model, although the arguments are also appropriate for situations where the $I(0)$ component may be a nonlinear formulation as suggested by Baillie and Kapetanios (2007). Sections 3 and 4 then provide some theoretical and simulation results on the properties of the parameter estimates from simple $ARFIMA$ models using alternative methods of inference. The first method is to estimate the long memory parameter from a SPE Local Whittle (LW) estimator, which is then used to fractionally filter the series before a second step estimation of the short run parameters. This method is referred to as the Local Whittle Two Step Estimator ($LWTSE$). The role of choice of bandwidth is also highlighted when assessing the properties of the $LWTSE$ methodology.

The second method is the simultaneous estimation of both the long memory and short memory parameters by MLE. The next section focus on the computation of the Impulse Response Weights ($IRWs$) when they are derived from the different estimation procedures of $LWTSE$, MLE and alternatively, through the use of a high order linear AR approximation. The use of an AR approximation is one important novelty of the current paper since such an approach, although discussed by Poskitt (2005), has never been advocated in the context of IRW analysis. The MLE is found to generally compare favorably with the LW approach. The high order AR approximation is found to perform surprisingly well; even when compared to MLE. This of course is extremely encouraging for the AR approximation since unlike MLE which is a parametric estimation method, the AR is nonparametric. Section 6 deals with some of the theoretical properties of the high order AR approximation approach and also discusses the applicability of the method in nonstationary environments. This is another useful innovation of the paper, as we find that high order AR approximations can deal seamlessly with both stationary and nonstationary long memory processes, for the purposes of IRW analysis.

The results generally cast considerable doubt on the purpose of using a SPE of d and the desirability of the current research focus in the literature. This paper argues that an investigator should give careful thought concerning the particular metrics that are important in a series, before automatically turning to using an SPE of the long memory parameter. In particular, the MLE, and surprisingly the high order AR approach have many desirable features. The paper ends with a short concluding section.

2 Models and Methods

A fractionally integrated, or long memory process, generates hyperbolic rates of decay in the autocorrelation function and impulse response weights of its realization. The simplest discrete time parameterization is the $ARFIMA$ model, which combines both long memory and short run $I(0)$ dynamics, and represent a flexible extension of the $ARIMA$ model. In particular, a univariate time series process, y_t, which is said to be fractionally integrated of order d, or $I(d)$

$$(1-L)^d y_t = u_t,\ t = 1, \ldots, T \tag{1}$$

where L is the lag operator and u_t is a short memory, $I(0)$ process. Then y_t is said to be a fractionally integrated process of order d, or $I(d)$. See Granger and Joyeux (1980), Granger (1980) and Hosking (1981) for further definitions and derivations of these processes. Note that the $I(0)$ process is defined as having partial sums that converge weakly to Brownian motion, while d represents the degree of "long memory", or persistence in the series. For $-0.5 < d < 0.5$ the process is stationary and invertible; while for $0.5 \leq d \leq 1$, the process does not have a finite variance, but still has a finite cumulative impulse response function. The Wold decomposition, or infinite order moving average representation of this process is given by

$$y_t = \sum_{i=0}^{\infty} \psi_k \epsilon_{t-k} \quad (2)$$

where $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2$, $E(\epsilon_t \epsilon_s) = 0$, $s \neq t$. For large lags k, these coefficients decay at the very slow hyperbolic rates of $\psi_k \sim c_1 k^{d-1}$ and similarly the infinite autoregressive representation coefficients decay at the rate of $c_2 k^{-d-1}$and autocorrelation coefficients at the rate of $c_3 k^{2d-1}$, where c_1, c_2 and c_3 are constants. The hyperbolic decay that is generated by such a process is known as the 'Hurst effect', after Hurst (1951), who first discovered the phenomenon in hydrological time series data.

If the short memory component is represented as a stationary and invertible $ARMA(p,q)$ process, then equation (1) becomes the well known $ARFIMA(p,d,q)$ model,

$$\phi(L)(1-L)^d y_t = \theta(L)\epsilon_t \quad (3)$$

where $\phi(L)$ and $\theta(L)$ are polynomials in the lag operator of orders p and q respectively, with all their roots lying outside the unit circle.

To some extent the model is deceptively simple and appears a relatively minor extension of the class of $ARIMA$ models analyzed by Box and Jenkins (1970) who developed subjective model selection procedures based on sample autocorrelations and sample partial autocorrelations of the regular, differenced and seasonally differenced, or transformed series. Although enormous strides have been made in the last twenty years concerning the appropriate testing for unit root processes and also the application of information criteria for model selection purposes, the basic strategy of Box and Jenkins (1970) remains valid today. Of particular significance is their emphasis on diagnostic testing or checking, and successive stages of model improvement[1] While the imposition of a unit root in an AR operator leads to a deterministic transformation of a time series, the parallel analogous plug in method of

[1]The methodological approach of Box and Jenkins (1970) is essentially based on the inductive approach of Karl Popper. Linear univariate $ARIMA$ models are typically specified from the concept of parsimony of parameters and specification testing is from simple models to possibly more complicated models, depending on the outcome of diagnostic testing. In contrast, linear $VARs$ seem easiest to specify from the general to specific criteria, while Clive Granger has indicated that most nonlinear time series models appear to require starting from specific models and to test outwards.

an SPE of the d parameter, immediately gives rise to a stochastic transformation of the series and far more complicated inferential problems for specification and estimation of short run dynamics, i.e. $ARMA$ structure. This leaves an open and essentially unanswered, the question as to how SPE of a long memory parameter can relate to simply constructing a model for a univariate time series process.

3 Parameter Estimation Issues

This section considers the use of the Local Whittle SPE to estimate the parameters of a univariate time series process. The procedure uses a two step estimation where an initial estimator of the long memory parameter d is obtained by a SPE such as the LW and the short memory process u_t in equation (1) is estimated from a fractionally filtering the original series y_t. The parameters of the short memory process u_t are then estimated from minimization of the conditional sum of squares function.

In general, the use of $SPEs$ can be motivated by the desire to avoid difficulties over the specification of the short run $ARMA$ parameters. There has been a long history of SPEs in this literature. Originally, Hurst (1951) developed the rescaled range statistic, denoted as R/S to detect long memory; which was adapted by Lo (1991) to produce the modified rescaled range test statistic. However, a very widely used early estimator has been the GPH log periodogram estimator of Geweke and Porter Hudak (1983). The performance of GPH is well known to be poor in the presence of substantial short run dynamics, e.g. see Agiakloglou, Newbold and Wohar (1992). Hence the generally preferred SPE is now the LW estimator; see Robinson (1995), Dalla, Giraitis, and Hidalgo (2005) and Phillips and Shimotsu (2006), and is therefore the method used in this study. The LW SPE for d, denoted by $\hat{d}_{LW}$ is obtained by minimizing the objective function

$$\ln\left[\frac{1}{m}\sum_{j=1}^{m}\omega_j^{2d}I(\omega_j)\right]-\frac{2d}{m}\sum_{j=1}^{m}\ln(\omega_j) \tag{4}$$

with respect to d, where $I(\omega_j)$ is the periodogram given by $I(\omega_j)=\frac{1}{2\pi T}\left|\sum_{j=1}^{T}y_te^{i\omega_jt}\right|^2$. The estimator depends on the choice of bandwidth, m and data dependent methods for bandwidth selection are discussed in Henry (2001). The LW estimator of d is known to have the limiting distribution of

$$m^{1/2}\left(\hat{d}_{LW}-d_0\right)\rightarrow N\{\mathbf{0},(\mathbf{1/4})\} \tag{5}$$

where d_0 denotes the true value of d, and m represents the choice of bandwidth in equation (4). It is important to note that $m \leq T^{4/5}$ and is generally chosen in the range of $T^{1/2} \leq m \leq T^{4/5}$ and in the usual case of ignorance of the short run dynamics, m is generally selected as $m = T^{0.5}$. However, when there is substantial persistence in the short run dynamics, the value of m should potentially be reduced so that more weight is placed on ordinates of the periodogram associated with the low frequency components.

For the process in equation (1) the spectral density function (s.d.f.) of y_t is denoted by $f(\omega)$, and the s.d.f. of u_t is $f^*(\omega)$. Hence $f(\omega) = |1 - \exp(i\omega)|^{-2d} f^*(\omega)$. The s.d.f. of can be approximated as $\omega \to 0+$, as $f(\omega) = \mathcal{L}(1/\omega)\omega^{-2d}\{1 + c\omega^{\beta} + o(\omega^{\beta})\}$ where $\mathcal{L}(1/\omega)$ is a slowly varying function $0 < c < \infty$, usually chosen as one, $\beta \in (0, 2]$ The issue of optimal bandwidth selection has been addressed by Henry (2001), who shows that the optimal bandwidth in the sense of minimizing MSE is given by m^*_{LW} and for the Local Whittle estimator is

$$m^*_{LW} = \left(\frac{3}{4\pi}\right)^{4/5} \left|\tau^* + \frac{d}{12}\right|^{-2/5} T^{4/5} \tag{6}$$

where

$$\tau^* = \left[\frac{f^{*''}(0)}{2f^*(0)}\right]_{\omega=0} \tag{7}$$

and τ^* has the interpretation of representing the degree of smoothness of the s.d.f. $f^*(\cdot)$ of the short memory component u_t, as the frequency approaches zero. Henry (2001) considers an iterative form of estimation which iterates between successive choices of $\hat{d}_{LW}$ and m^*_{LW}. However, this study uses the optimal bandwidth in a different context and calculates the theoretical optimum bandwidth m^*_{LW} given knowledge of the underlying theoretical data generating process. This ensures the maximum advantage of the $LWTSE$ methodology when comparing it with the alternatives of MLE and high order autoregressive approximation methods. The choice of m^*_{LW} is described in more detail for particular short memory processes in the context of the Monte Carlo experiments in section 4 of this paper.

The next step given the Local Whittle estimate $\hat{d}_{LW}$ of the long memory parameter, is to consider the estimation of the short memory parameters. First, suppose d is known; then the observed y_t series can be fractionally filtered to obtain

$$u_t = y_t - \sum_{l=1}^{t-p} \pi_l(d) y_{t-l} \tag{8}$$

where $(1-L)^d y_t = y_t - \sum_{l=1}^{\infty} \pi_l(d) y_{t-l}$, and $\pi_l(d)$ are the coefficients of the infinite AR representation of y_t in terms of u_t, so that

$$\pi_l(d) = \Gamma(l-d)\Gamma(-d)^{-1}\Gamma(l+1) \tag{9}$$

In practice, d is unknown and has to be replaced by its estimate $\hat{d}_{LW}$. Then, the Feasible Fractionally Filtered (FFF) series based on observables is

$$\hat{u}_t = y_t - \sum_{l=1}^{t-p} \hat{\pi}_l(\hat{d}_{LW}) y_{t-l} \tag{10}$$

where $\hat{\pi}_l(\hat{d}_{LW}) = \Gamma(l-\hat{d}_{LW})\Gamma(-\hat{d}_{LW})^{-1}\Gamma(l+1)$. For concreteness, this section of the paper will focus on the estimation of the univariate $ARFIMA(p,d,q)$ process. Therefore, y_t is assumed to be generated by the $ARFIMA(p,d,q)$ process in (3). The complete parameter vector is denoted by $\vartheta = (d,\beta)'$, where the $(p+q)$ $ARMA$ parameters are in the vector $\beta = (\phi_1, ..\phi_{\mathbf{p}}, \theta_1, ...\theta_q)'$. Let $\beta_0(d_0)$ denote the true parameter values, and let $\hat{\beta}(\hat{d}_{LW})$ be the $LWTSE$ of β based on the feasible fractionally filtered series $\hat{u}_t$. Hence the $ARMA(p,q)$ parameters of the original $ARFIMA(p,d,q)$ process in equation (2) are estimated by minimizing the conditional sum of squares, CSS, conditional on $\hat{d}_{LW}$. The following result provides consistency and a rate of convergence for the two step estimator of the $ARFIMA(p,d,q)$ model. The result is proven in the appendix.

Theorem 1 *Let y_t be given by the $ARFIMA(p,d,q)$ process in (3), where $\phi(L)$ and $\theta(L)$ are polynomials in the lag operator of orders p and q respectively, with all their roots lying outside the unit circle; and the disturbance ϵ_t is i.i.d.$(0,\sigma^2)$, with $E(\epsilon_t^4) < \infty$. Then,*

$$\hat{\beta}(\hat{d}_{LW}) - \beta_0(d_0) = O_p(m^{-1/2})$$

The above result provides some necessary equipment for examining the properties of the use of the $LWTSE$.[2] In order to provide a valid benchmark for comparison of the $LWTSE$, it is necessary to compare the methodology with the obvious alternative of MLE. On assuming the innovations in the $ARFIMA(p,d,q)$ process in equation (3) to be $\epsilon_t \tilde{\ } NID(0,\sigma^2)$, then the Gaussian likelihood is numerically

[2] The investigation of the properties of two step estimators based on consistent first step estimators is a long standing issue in econometrics and an early treatment of related issues can be found in Amemiya (1973). The problem in the current paper is slightly different to other time series econometrics applications, since an estimator with a slower rate of convergence is being used in the first step.

maximized with respect to the complete vector of parameters ϑ. Under these conditions, Fox and Taqqu (1986) have shown the asymptotic distribution of the MLE to be

$$T^{1/2}\left(\hat{\vartheta}-\vartheta_0\right) \rightarrow N\{\mathbf{0}, \mathbf{A}(\vartheta_0)^{-1}\}, \tag{11}$$

where ϑ_0 denotes the true value of the vector of parameters, and where $\mathbf{A}(\vartheta_0)$ is the information matrix. The results of Fox and Taqqu (1986) are obtained by considering the demeaned process, while Li and McLeod (1986) has shown that $T^{1/2}$ consistency and asymptotic normality are achieved when the unconditional mean is zero or known. Dahlhaus (1989) and Moehring (1990) have shown that the same asymptotic properties hold also when the unconditional mean is not known and have to be estimated. In this case the MLE of the parameter estimates will be $T^{1/2}$ consistent. The inclusion of an intercept parameter, will result in a $T^{1/2-d}$ consistent estimator. An especially important result has been achieved by Hosoya (1997) who showed that

$$T^{1/2}\left(\hat{\vartheta}-\vartheta_0\right) \rightarrow N\{\mathbf{0}, \mathbf{A}(\vartheta_0)^{-1}B(\vartheta_0)\mathbf{A}(\vartheta_0)^{-1}\}, \tag{12}$$

where $\mathbf{A}(.)$ is the Hessian and $\mathbf{B}(.)$ is the outer product gradient, both of which are evaluated at the true parameter values ϑ_0. This result is valid when the Gaussianity assumption is relaxed and replaced with the assumption that the innovations in (3) merely satisfy some mild mixing conditions. Given the Hosoya (1997) result, the implementation of quasi MLE is then straightforward. Hosoya (1997) provides results for more general long memory models than $ARFIMA$ models while allowing the relaxation of the Gaussianity assumption. More recently Baillie and Kapetanios (2008) have extended this analysis to include the estimation of long memory models with nonlinear autogressive structures and have found that $QMLE$ also works well in these circumstances. The computation of such models is aided by the minimization of the conditional sum of squares (CSS). This estimator has been shown by Robinson (2006) to be a consistent and asymptotically normal and to be approximately equivalent to the Gaussian MLE.

4 Simulation Results for Estimators

The small sample properties from using the $LWTSE$ of d and the short memory parameters from the FFF series are analyzed in the following simulation study.

In order to provide a valid benchmark, corresponding results from MLE were also computed. In the simulation experiments, realizations of $ARFIMA(1, d, 0)$ processes were generated for the three different sample sizes of $T = 200$, $T = 400$ and $T = 1,000$ and for 24 different simulation designs. These experiments were in blocks of four different values of the AR coefficient $\phi = 0.50$, $\phi = 0.90$, $\phi = 0.95$, and $\phi = 0.99$. For each set of the four different autoregressive parameter values, there are six different values of the long memory parameter, namely $d = 0.0$, $d = 0.10$, $d = 0.20$, $d = 0.30$, $d = 0.40$ and $d = 0.49$.

The biases and variances of the MLE and the $LWTSE$ of the long memory parameter, d and the autoregressive parameter ϕ, are compared across the different designs and sample sizes. The MLE are denoted by $\hat{d}_{MLE}$ and $\hat{\phi}_{MLE}$; while the $LWTSE$ consist of the LW estimate of the long memory parameter from the original observed series and is denoted by $\hat{d}_{LW}$ and the autoregressive parameter estimated in the second step from the FFF series and is denoted by $\hat{\phi}_{LWTSE}$.

Table 1 presents the results concerning the different methods for estimation of the long memory parameter, d. The first three panels present the biases of the parameter estimates and the last three panels are concerned with their corresponding variances. The first and fourth panels refer to the bias and variance respectively of , $\hat{d}_{LW}$ when it is computed from the widely recommended and heavily used bandwidth of $m = T^{1/2}$, since this is the usually recommended value in the absence of any prior information. In fact, this is the low end of the generally recommended range of $T^{0.5} \leq m \leq T^{0.8}$. The second and fourth panels of the table 1 refer to the bias and variance of the $\hat{d}_{LW}$ when it is computed from the optimal bandwidth discussed earlier. It should be noted again that this analysis provides the most favorable circumstances since the Monte Carlo work is performed from implementing the known values of the short memory, autoregressive parameter, which would not be known in practice when the bandwidth was chosen. Hence these results can be legitimately interpreted as an upper bound on the performance of the LW estimator, $\hat{d}_{LW}$ and that of the subsequent $LWTSE$ of ϕ, i.e. $\hat{\phi}_{LWTSE}$. For the $ARFIMA(1, d, 0)$ process, it can be shown that τ^* in equation (7) reduces to $\tau^* = \frac{-\phi}{(1-\phi)^2}$ and the optimal bandwidth m^*_{LW} can be determined from equation (7) for each sample size T and for each design in the parameter space. For example, when $T = 1,000$, the optimal bandwidths are presented in Table 3 in the appendix. Finally, the third and sixth panels of Table 1 present the bias and variance of the MLE, of d denoted by $\hat{d}_{MLE}$ when there is joint estimation of d and ϕ.

The layout and presentation of the Table 2 is analogous to that of Table 1, except that it deals with the bias and variance properties of the estimated autoregressive parameter. The first and fourth panels correspond to the $\hat{\phi}_{LWTSE}$ when it is estimated from the FFF series derived from $\hat{d}_{LW}$ when a bandwidth of $m = T^{1/2}$ has been used. The second and fifth panels are analogous and only differ in the the fact that they are based on the optimal bandwidth. While the third and sixth panels are relevant for the MLE of ϕ when it has been simultaneously estimated with d.

With relatively little persistence in the autocorrelation of the short memory component, i.e. $\phi = 0.5$, all the estimation methods including the $LWTSE$ perform very well. However, the most striking aspect of tables 1 and 2 are the very poor performance of the $LWTSE$ when there is a reasonable level of persistence in the short memory process (i.e. for $\phi = 0.90$ or more). In particular, on first focusing on the $LWTSE$ for the $m = T^{1/2}$ bandwidth, it is seen from the first and fourth panels of tables 1 and 2 that there are very substantial upward biases in the estimate $\hat{d}_{LW}$ and a severe corresponding downward bias in the estimator $\hat{\phi}_{LWTSE}$. For example, when the true data generating process is an $ARFIMA(1, d, 0)$ process with $d = 0.30$ and $\phi = 0.90$, the average estimated values from employing the $LWTSE$ are $\hat{d}_{LW} = 0.56$ and $\hat{\phi}_{LWTSE} = 0.70$. These compensating biases get worse with increasing persistence, so that for example with a data generating process of an $ARFIMA(1, d, 0)$, with $d = 0.30$ and $\phi = 0.95$, the average estimated values from the $LWTSE$ are $\hat{d}_{LW} = 0.79$ and $\hat{\phi}_{LWTSE} = 0.54$. In general, the magnitude of the bias and variability of the $LWTSE$ are very disappointing and the methodology appears unable to discriminate between the short run $I(0)$ autoregressive dynamics and long memory $I(d)$ dynamics.[3]

Moving on to the optimal bandwidth, several features are apparent. First, the previous corresponding trade off in terms of bias of the estimates of d and ϕ when $m = T^{0.5}$ is no longer present. Second, although the properties of the $\hat{d}_{LW}$ are considerably improved in terms of bias, there is a corresponding increase in parameter estimation variability due to the considerably smaller value of m and hence reduced rate of consistency through equation (5). Furthermore, this

[3] The disappointing small sample performance of Local Whittle has been documented in the literature by Nielsen and Frederiksen (2004), but only in the context of estimating d. No work on the small sample properties of estimation strategies for short memory dynamics is available to the best of our knowledge. As argued in the introduction, adequate modelling of the short memory dynamics is of paramount importance for most modelling purposes such as IRW analysis and forecasting.

increased variability is passed on to the FFF series $\hat{u}_t$ second stage estimation of the ϕ parameter is not generally improved. This is generally due to the $\hat{d}_{LW}$ estimate being based on a very small number of ordinates, m which increases the sampling variability of the FFF series $\hat{u}_t$ and hence to the autoregressive parameter estimate $\hat{\phi}_{LWTSE}$, which has a reduced bias compared to before, but much larger variability.

The most relevant benchmark is the estimation of the model simultaneously by MLE and these results are presented in the third and sixth panels of tables 1 and 2. In contrast, the biases of the $MLEs$ for both ϕ and d are quite small for all values of (d, ϕ) except for $\phi = 0.99$ and $d = 0.40$, and $\phi = 0.99$ and $d = 0.49$. Clearly the sampling distribution of the MLE appears to be breaking down when there is the simultaneous move to non stationarity of both the autoregressive and long memory parameters. Otherwise, MLE appears to perform extremely well; and similar remarks are also valid for the variances of the MLE. These results are indicative of considerable gains from using MLE compared with $LWTSE$ especially when the short memory process is strongly autocorrelated.

Several other simulation experiments were also conducted which are not fully reported here in order to conserve space, but are available from the authors on request. In particular, the most important aspect of the short memory component was found to be the persistence it exhibited, while relatively low order moving average processes were not a particular problem for the $LWTSE$. The combination of persistence and cycles due to the presence of complex conjugate roots in the autoregressive operator was found to be particularly problematic for the $LWTSE$[4] For the $ARFIMA(2, d, 0)$ process

$$(1 - L)^d (1 - \phi_1 L - \phi_2 L^2) y_t = \varepsilon_t \tag{13}$$

it can be shown after some algebra that

$$\tau^* = \frac{(\phi_1 \phi_2 - \phi_1 - 4\phi_2)}{(1 - \phi_1 - \phi_2)^2} \tag{14}$$

For example, the process $(1 - L)^d (1 - 0.0800L - 0.9025L^2) y_t = \varepsilon_t$ has an autoregressive component with complex roots which has a damping factor of $|-\phi_2|^{1/2} = 0.98$, which generates very slow decay it its autocorrelation function and impulse response weights with a sinusoidal pattern superimposed. The results for the $LWTSE$ are particularly poor compared with the MLE in table 4, where for

[4] We are very grateful to Peter Phillips for this insight.

brevity are focused on the large sample size of $T = 5,000$ and a bandwidth of $m = [T^{0.5}] = [(5,000)^{0.5}] = 71$. The results for the MLE for the designs of $d = 0.0, 0.10, 0.20, 0.30$ and $d = 0.40$ generally indicate relatively small biases for all the estimates of d, ϕ_1 and ϕ_2. For the design with $d = 0.49$, the MLE of d has a very slight increase in bias, while there are compensating biases in the estimates of ϕ_1 and ϕ_2, with the bias in the sum of ϕ_1 and ϕ_2 being negligible. Hence the total persistence, as measured by the damping factor, is quite well estimated by MLE. In contrast, the LW of d has substantial bias across all the $ARFIMA(2, d, 0)$ designs, with the bias increasing as d increases. Clearly, even in this very large sample size, the strong persistence in the $AR(2)$ short memory component causes significant problems for the LW of d and also for the $LWTSE$ of the two autoregressive parameters. Similar results were also computed for the optimal bandwidth which was $m = 7$ across all the range of values of d being examined. Although the bias of the estimate of d was considerably reduced, the very small bandwidth created difficulties with the second step estimates of the short memory parameters.

Finally, it should be noted that the above results are all predicated on the assumption that the order of the short memory component, i.e. an $AR(1)$, or $AR(2)$ process is known and that the only unknown quantities are the parameter values. This assumption has been applied to both the $LWTSE$ and the MLE and in this sense treats both methods equally. In practice the order of say an $ARMA(p, q)$ process for u_t would have to be determined before the parameter estimation method is implemented. In order to increase the realism of the results a limited simulation was conducted where the order p of an autoregressive process was determined by information criteria for an $ARFIMA(p, d, 0)$ model when using MLE and also on the FFF series $\hat{u}_t$ when using the $LWTSE$ approach. While this simulation slightly worsened both methods of estimation it did not lead to any significant change in the relative performance of the MLE and $LWTSE$ methods.

5 Impulse Response Weight Analysis

As argued in the introduction, the previous literature on long memory processes seems to have focused too much on the properties of one parameter estimate and not sufficiently on the general interpretation and usefulness of the whole model. Also, the previous section of this paper has presented evidence on the relatively poor performance of the $LWTSE$ in terms of parameter estimation. However, it seems equally important to assess the overall characteristics and features of the model. A widely used and economically important concept are the implied Impulse Response Weights ($IRWs$) from an estimated model. Indeed, in many practical situations an investigator will wish to estimate a time series model such as $ARFIMA$ for the specific purpose of IRW analysis, or for making short to medium term predictions.

The IRW analysis has been adopted in several applied econometric studies which have used long memory models. In particular, the persistence of real GNP growth and its relationship to money growth has been investigated by Diebold and Rudebusch (1989), who use the frequency domain MLE of Fox and Taqqu (1986), and also Sowell (1992b)'s time domain MLE. A multivariate analysis of $IRWs$ was provided by Jensen (2007) who used time domain MLE to estimate the parameters in a vector $ARFIMA$ model which includes output and money growth rates.

Many articles have used long memory models to analyze the rates of persistence in real exchange rates and the duration of deviations from purchasing power parity; in particular see Diebold, Husted, and Rush (1991), who use the frequency domain MLE of Fox and Taqqu (1986). Other examples include the study by Baillie, Chung and Tielsau (1996), who use a time domain MLE for a study of inflation and the Friedman hypothesis. While Andersen, Bollerslev, Diebold and Labys (2004) use a multivariate Local Whittle estimator for the Realized Volatility of nominal exchange rates and a subsequent vector process to generate the implied $IRWs$. Andersen, Bollerslev, Diebold and Labys (2004) then essentially apply the FFF to Realized Volatility of currency returns and then estimate a VAR on the filtered series. It is interesting to note that Diebold and Rudebusch (1989) discuss the possibility of using a SPE such as GPH, but decide not to on the grounds that it is difficult to recover the correct standard errors of the estimated $ARMA$ parameters.[5]

In order to assess the relative importance of the estimation method for IRW analysis, the data generating process was assumed to be the $ARFIMA$ process in equation (2). The implied $IRWs$ are denoted by ψ_k for $k = 1,,2,....$ and are generated from the representation $\psi(L) = \theta(L)(1-L)^{-d}\phi(L)^{-1}$ where $\psi(L) =$

[5]Similar problems have emerged on IRW of the conditional variances of long memory $GARCH$ models. However, this is a separate topic and is beyond the scope of the current paper

$\sum_{k=1}^{\infty} \psi_k L^k$ and the underlying "true" model is the $ARFIMA(p, d, q)$ process in equation (3). The practical calculation of the $IRWs$ requires replacing the true theoretical parameters with their corresponding estimates. When using the $LWTSE$, the long memory parameter d is replaced with $\hat{d}_{LW}$ and the $ARMA$ parameters are replaced by the minimized CSS estimates from the filtered series $\hat{u}_t$. In the case of the $ARFIMA(1, d, 0)$ model, this clearly amounts to using the quantities $\hat{d}_{LW}$ and $\hat{\phi}_{LWTSE}$. The efficacy of this approach was investigated with the following simulation study which was based on the same designs as described earlier in section 4 in the context of parameter estimation.

The $IRWs$ derived from the true model specification with known parameter values were plotted as a *solid line* for the first forty lags in the figures 1 through 4. As noted earlier for large lag k, these Wold decomposition coefficients decay at the approximate rate of $\psi_k \sim c_1 k^{d-1}$ for large lags k. However, the presence of a relatively persistent $AR(1)$ component process can considerably delay and change the appearance of the $IRWs$ for short to moderate lag lengths. These figures report IRW for the case where d takes the values $0.3, 0.49$ and the short memory, autoregressive parameter takes the values 0.5 and 0.95.

When the $ARFIMA$ parameters are replaced by their corresponding $LWTSE$ the estimated $IRWs$ can be calculated and are denoted in the figures by the long dashes "– –". Again, in all these cases the bandwidth of $m = T^{0.5}$ was used. The optimal bandwidth results were comparable to those reported here. When the $ARFIMA$ parameters are replaced by their corresponding MLE the estimated $IRWs$ are denoted by the medium length dashes "- -". Figures 1 through 4 report these different $IRWs$ for lags $k = 1, 2, ...40$.

For low levels of autocorrelation in the short run dynamics, i.e. with $\phi_1 = 0.5$, all methods perform well and as can be seen from Figures 1 and 3, there is virtually no difference between them at the larger sample size of $T = 1,000$. However, in the presence of fairly persistent short run dynamics autocorrelation, substantial differences between the different methods begins to emerge. In general the $IRWs$ generated from the MLE performs far better than those derived from the $LWTSE$. For many situations the differences in performance between the two methods is quite marked with the general superiority of the benchmark MLE method. However, for the design $\phi = 0.95$ and $d = 0.49$ both the MLE and the $LWTSE$ perform very poorly in terms of closeness to the true $IRWs$. Hence these results also indicate the potential problems arising from estimating IRWs based on the $LWTSE$. Many of the issues with regard to the quality of parameter estimation are also apparent in the simulations of the IRW analysis.

6 An Autoregressive Approximation Method

Some suggestions in the literature have recommended fitting a high order $AR(p)$ process to the FFF series after having first used Local Whittle estimation. However, this naturally leads to the question as to why the Local Whittle estimate is worth using at all, since the estimation of $ARFIMA(P, d, 0)$ models is most easily accomplished by a one step MLE and furthermore this procedure can be seen from the above analysis to generally provide superior estimates of the model's parameters and also of the $IRWs$.

An even more straightforward approach would be to implicitly ignore the presence of long memory and to simply estimate a high order $AR(P)$ model, which is a truncation of the infinite order autoregressive expansion in equation (2). Such a a model would then be used for forecasting and generation of $IRWs$. Since the $ARFIMA(p, d, q)$ model can be represented by an infinite autoregressive expansion of the form

$$y_t = \sum_{j=1}^{\infty} \pi_j y_{t-j} + v_t \tag{15}$$

a possible method would be to directly estimate by OLS the truncated autoregressive, $AR(P)$, expansion

$$y_t = \sum_{j=1}^{P} \pi_j y_{t-j} + \tilde{v}_t \tag{16}$$

where the order P, is obtained by some information criterion. This approach has recently been theoretically analyzed by Poskitt (2005). In particular, Poskitt (2005) has shown some useful theoretical results. On denoting the least squares estimates of π_j, obtained by fitting an $AR(P)$ model to the data, by $\hat{\pi}(j)^{(P)}$; and on further denoting the coefficients that solve the Yule-Walker equations for an $AR(P)$ model by $\pi(j)^{(P)}$, then theorem 5.1 of Poskitt (2005) states that $\sum_{j=1}^{P} \left|\hat{\pi}(j)^{(P)} - \pi(j)^{(P)}\right|^2 = o_p(1)$ for all P such that $P \to \infty$ and $P = o(T^{\alpha})$ for all $\alpha > 0$. For example, an acceptable sequence for P is $(lnT)^{\alpha}$ for some $\alpha > 1$. Further, by the extension of Baxter's inequality proven in Theorem 4.1 of Inoue and Kasahara (2006) it follows that$\sum_{j=1}^{P} \left|\pi(j)^{(P)} - \pi(j)\right| = o(1)$, as long as $P \to \infty$.

Then, overall,$\sum_{j=1}^{P} \left|\hat{\pi}(j)^{(P)} - \pi(j)\right|^2 = o_p(1)$ which implies that the $IRWs$ can be consistently estimated by fitting an approximating autoregressive model to the long memory time series realization. A major remaining issue concerns the choice of P. Poskitt (2005) has shown, via his Theorem 5.3, that selecting P by information criteria such as the AIC or BIC is asymptotically efficient in the sense of Shibata (1980). In the Monte Carlo study in this paper the value of P is fixed at $(lnT)^2$.

A further attraction of using an infinite $AR(P)$ approximation for a long memory process is that the validity of the approach extends to non stationary processes. Non-stationary long memory process are still amenable to impulse response analysis, since when $0.5 \leq d \leq 1$ the process does not have a finite variance, but still has finite cumulative $IRWs$. Hence shocks to a long memory process with $0 < d \leq 1$ are transient unlike the random walk case when $d = 1$ where the shocks are permanent. An extension to non-stationary long memory processes is provided by the following theorem which is proven in the appendix

Theorem 2 *Consider the infinite autoregressive process* y_t *in (15) where* $\pi_j = O(j^{-d-1})$ *for large* j, $0.5 < d < 1$. v_t *is i.i.d. with finite second moments. Assume that this AR process is invertible and can be equivalently represented as* $y_t = \sum_{j=1}^{\infty} \psi_j v_{t-j}$ *where* $\psi_j = O(j^{d-1})$ *for large* j. *It is further assumed that the sample second moment of the regressors in (16), normalized by* T^{-2d} *has asymptotically non-zero and bounded eigenvalues in probability. Further,* $P = o\left(T^c\right)$ *for all* $c > 0$. *Then,*

$$\left(\sum_{j=1}^{P} \left|\hat{\pi}(j)^{(P)} - \pi(j)\right|^2\right)^{1/2} = o_p(P).$$

The above is a consistency result in the sense that inconsistency of each $\hat{\pi}_j$ would imply $\|\hat{\pi} - \pi\| = O_p(P)$. A further important aspect of the above Theorem is that for all elements in the set $\{1, 2, ..., P\}$ but a set of cardinality P_c such that $P_c/P \to 0$, then $|\hat{\pi}_j - \pi_j| = o_p(1)$. Furthermore, the above result fits neatly with the result obtained by Poskitt (2005) whereby for $d < 1/2$

$$\sum_{j=1}^{P} \left|\hat{\pi}(j)^{(P)} - \pi(j)\right|^2 = O_p\left(P \frac{\ln T}{T^{1-2d}}\right) \tag{17}$$

implying that for values of d close to 0.5, then $\|\hat{\pi} - \pi\| = o_p(P)$, which is identical to the above theorem . Interestingly, the value $d = 1/2$ is a worst case scenario since by (22) it can be seen that the rate improves as d increases beyond 0.5. This very interesting property of the high order AR approach makes the methodology attractive to use in a non stationary environment. This motivates a comparison of LW, the AR approximation, and, as a benchmark, MLE, for the estimation of IRWs in nonstationary settings. It should be noted that there are some existing results in the literature concerning the theoretical properties of LW and MLE when applied to nonstationary long memory processes. In particular, Velasco (1999) and Phillips and Shimotsu (2004) have shown that the LW estimator of d is consistent when $0.5 < d < 1$, and asymptotically normally distributed when $d < 3/4$, but not otherwise. For parametric estimation, there is work by Tanaka (1999), Ling and Li (2001) and Johansen and Nielsen (2008). The study by (2008) is especially relevant since it proves consistency of MLE for a parametric autoregressive nonstationary long memory model. The model is related but not equivalent to a nonstationary $ARFIMA$ model.

Given the above results, this study now considers Monte Carlo evidence for the $IRWs$ for non-stationary data generating processes, using the three estimation methods previously described. The estimated $IRWs$ are derived for the three different designs of ($d = 0.6$, $\phi = 0.95$), ($d = 0.8$, $\phi = 0.5$) and ($d = 0.8$, $\phi = 0.95$) and are graphically illustrated in Figures 5 through 7 respectively. The representation of the $IRWs$ derived from the different estimation methods is identical to that of the stationary processes considered earlier and represented in Figures 1 through 4. Some interesting stylized facts emerge from analysis of the results for the large sample size of $T = 1,000$. For the design ($d = 0.8$, $\phi = 0.5$) in figure 6, the MLE performs extremely well, with the high order AR approximation generally being slightly superior to the $LWTSE$. However, for the designs of ($d = 0.6$, $\phi = 0.95$) and ($d = 0.8$, $\phi = 0.95$) in figures 5 and 7 respectively, the high order AR approximation performs outstandingly well, with the MLE a poor third compared with the $LWTSE$. Hence there seems some evidence that MLE works well for non stationary long memory processes provided that there is only moderate degree of persistence in the short run dynamics. However, when the phenomenon of a non stationary long memory processes is combined with very persistent short run component the high order AR approximation method is extraordinarily accurate compared with MLE and the $LWTSE$.

To summarize, the results from using the high order $AR(p)$ model are unexpectedly good, since the method generally performs very well even when compared with the $MLEs$ being used to generate the $IRWs$. This exceptional result holds for both stationary and nonstationary processes. Given the excellent performance of this method, it raises important questions as to whether it is worth an investigator being concerned with the presence of long memory if the investigator's main interest is to only to assess the impact of shocks or innovations on a series.

7 Conclusions

This paper has addressed the issue concerning the applicability of the SPE Local Whittle (LW) estimator of the long memory parameter in a univariate time series. In virtually all practical applications it is expected that an investigator would typically want to estimate a complete model involving both short and long memory features, and also to generate Impulse Response Weights (IRWs) and predictions at a variety of horizons. Hence the paper analyzes the properties of the Local Whittle Two Step Estimator ($LWTSE$) where subsequent estimates of short run parameters are obtained from a fractionally filtered series. The $LWTSE$ is shown to generally have quite poor properties in the presence of moderately persistent autocorrelation in the short run component. The selection of an "optimal" bandwidth is generally not very helpful and either way the LW has either severe bias and/or very high sampling variability due to the rate of convergence being equivalent to the bandwidth selection.

The above difficulties with the LW estimation of the long memory parameter are also inherited by the estimated IRWs. The alternative approach of using a time domain MLE to jointly estimate the parameters of the model and subsequently the IRWs is shown to be generally superior to using the $LWTSE$. An alternative purely non parametric approach is to use a high order linear autoregression (AR) approximation. This procedure is shown to provide consistent estimates of the infinite autoregressive approximation. A very attractive feature of the AR method is that it is valid for non stationary regions of the parameter space. In general high order AR method appears to provide relatively good estimates of IRW both for stationary and nonstationary processes. It should be emphasized that unlike both the $LWTSE$ and MLE, the AR approximation is nonparametric for both the long and short memory parts of the process. This considerably increases the attractiveness of the AR procedure.

Finally the results in this paper also appear very relevant in a wider methodological context of how best to deal with time series that exhibit long memory properties. The high order AR approximation method is highly desirable if an in investigator wishes to obtain estimates of the IRWs and also especially in the case when there is considerable uncertainty about the stationarity of the process. If an investigator's primary purpose is to model the univariate time series, then using the $LWTSE$ is fraught with problems whenever there is moderately persistent autocorrelation in the short memory component. In this situation careful use of model selection criteria and application of $QMLE$ will probably be the best strategy. In more extensive empirical research an important preliminary issue is often to obtain an estimate of each variable's degree of persistence before embarking on some modeling exercise. Once again the problems associated with SPE and LW would make one skeptical of relying unduly on these types of estimators. The approach of jointly modeling short and long memory components would appear to again be the most fruitful approach.

Appendix

Proof of Theorem 1

Since all the roots of the polynomials in the lag operator $\phi(L)$ and $\theta(L)$ lie outside the unit circle, it follows that $\sum_{k=0}^{\infty} \pi_k^2 < \infty$ and hence that $\sum_{k=1}^{t-1} \pi_k y_{t-k} = O_p(1)$. The Local Whittle estimator $\hat{d}_{LW}$ will generate the FFF series $\hat{u}_t = (1-L)^{\hat{d}_{LW}} y_t = y_t - \sum_{l=1}^{t-p} \hat{\pi}_l(\hat{d}_{LW}) y_{t-l}$, where $\hat{\pi}_l(\hat{d}_{LW}) = \Gamma(l - \hat{d}_{LW})\Gamma(-\hat{d}_{LW})^{-1}\Gamma(l+1)$. Since $\hat{u}_t = (1-L)^{\hat{d}_{LW}} y_t$ then $(\hat{u}_t - u_t) = \sum_{j=1}^{\infty} \pi_j(\hat{d}_{LW} - d_0) u_{t-j}$. Since $(\hat{d}_{LW} - d_0) = O_p(m^{-1/2})$ and $u_t = (1-L)^d y_t$, then following the same approach as Wright (1995),

$$T^{-1} \sum_{j=1}^{\infty} (\hat{u}_t - u_t)^2 = T^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{t-1} \pi_j(\hat{d}_{LW} - d_0) u_{t-j} \right)^2$$

Then, using the mean value theorem we have that $\pi_j(d) = dX_j^1 + d^2 X_j^2$, where X_j^1 denotes the first derivative and X_j^2 the second derivative of $\pi_j(.)$. Then

$$\sum_{k=1}^{t-1} \pi_k u_{t-k} = d \sum_{k=1}^{t-1} X_j^1 u_{t-j} + d^2 \sum_{k=1}^{t-1} X_j^2 u_{t-j}$$

and following the same arguments as in Wright (1995), $(\hat{d}_{LW} - d_0) \sum_{k=1}^{t-1} X_j^1 u_{t-j} = O_p(m^{-1/2})$, and $T^{-1}(\hat{d}_{LW} - d_0) \sum_{k=1}^{t-1} X_j^2 u_{t-j} = O_p(m^{-1/2})$, and hence

$$T^{-1}\sum_{t=1}^{T-k}\hat{u}_t\hat{u}_{t+k} = T^{-1}\sum_{t=1}^{T-k}u_t u_{t+k} + O_p(m^{-1/2}) \tag{18}$$

This suffices to prove the Theorem result for an $ARFIMA(p,d,0)$ model. For the general case of an $ARFIMA(p,d,q)$ model we have that for the second step $ARMA$ estimation, the conditional MLE needs to be numerically maximized. Let us denote the likelihood function by $L(\beta)$. The form of the likelihood may be found in, e.g., (5.6.3) of Hamilton (1994). It is then clear that the likelihood function is differentiable and as long as (18) holds we have $L\left(\hat{\beta}(\hat{d}_{LW})\right) - L\left(\beta_0(d_0)\right) = O_p(m^{-1/2})$. But, by an application of the mean value theorem we have that $L\left(\hat{\beta}(\hat{d}_{LW})\right) = L\left(\beta_0(d_0)\right) + \frac{\partial L}{\partial \beta}\Big|_{\beta=\bar{\beta}}\left(\hat{\beta}(\hat{d}_{LW}) - \beta_0(d_0)\right)$. Hence, the result of the Theorem holds for $ARFIMA(p,d,q)$ models completing the proof.

Proof of Theorem 2

Let the norm of a matrix A, $\|A\|$, be given by $tr(A'A)^{1/2}$. The Theorem is proven if we show that $\|\hat{\pi}-\pi\| = o_p(P)$ where $\pi = (\pi_1, ..., \pi_P)$ and $\hat{\pi} = (\hat{\pi}_1, ..., \hat{\pi}_P)$. We have that $\|\hat{\pi}-\pi\| = \left\|(X'X)^{-1}X'\tilde{v}\right\|$, where $X = (y_1, ..., y_P)$, $y_i = (y_{P-i+1}, ..., y_{T-i})'$ and $\tilde{v} = (\tilde{v}_{P+1}, .., \tilde{v}_T)'$. Then,

$$\left\|(X'X)^{-1}X'\tilde{v}\right\| \leq \left\|(X'X)^{-1}\right\| \|X'\tilde{v}\| \tag{19}$$

We examine each term of the right hand side of (19) in turn. It is possible to first determine the order of magnitude of each diagonal term of $(X'X)^{-1}$. We consider Kantorovich's inequality for a square $m \times m$ matrix A, which states that for the i,i-th element of A^{-1} the following holds: $(A^{-1})_{ii} \leq \frac{1}{4a_{ii}}\left(\frac{\alpha}{\beta}+\frac{\beta}{\alpha}+2\right)$, where a_{ii} is the i,i-th element of A, $\alpha = \min_i \lambda_i(A)$, $\beta = \max_i \lambda_i(A)$ and λ_i denotes the i-th eigenvalue of A. By the positive definiteness of the matrix and the assumption of the Theorem it is known that all eigenvalues of $X'X$ are of the same order of magnitude. Hence, $\left(\frac{\alpha}{\beta}+\frac{\beta}{\alpha}+2\right)$ is $O_p(1)$ in this case. Further, by the positive definiteness of the matrix we have that $\left|((X'X)^{-1})_{ij}\right| \leq \left|((X'X)^{-1})_{ii}\right|$ when $i \neq j$. As a result, we focus on the behavior of $\sum_{t=P+1}^{T} y_t^2$. By Beran (1994),

$$\sum_{t=P+1}^{T} y_t^2 = O_p(T^{2d}) \tag{20}$$

Hence, $\left|((X'X)^{-1})_{ii}\right| = O_p(T^{-2d})$. It then easily follows that $\left\|(X'X)^{-1}\right\| = O_p(PT^{-2d})$. Moving on to $\|X'\tilde{v}\|$ we need to determine the behavior of $\sum_{t=P+1}^{T} y_{t-i}\tilde{v}_t$.

We have

$$\sum_{t=P+1}^{T} y_{t-i}\tilde{v}_t = \sum_{t=P+1}^{T} y_{t-i}\left(\sum_{j=P+1}^{t} \pi_j y_{t-j} + v_t\right) \leq \sum_{j=P+1}^{T} \pi_j \left(\sum_{t=P+1}^{T} y_{t-i}y_{t-j}\right) + \sum_{t=P+1}^{T} y_{t-i}v_t.$$

Then, it is necessary to show that

$$\sum_{t=P+1}^{T} y_{t-i}v_t = O_p\left(T\right) \tag{21}$$

But, by the independence of v_t it follows that $y_{t-i}v_t$ is a martingale difference sequence and so (21) follows by the martingale difference law of large numbers. By (20), $\sum_{t=P+1}^{T} y_{t-i}y_{t-j} = O_p(T^{2d})$. Since, $\pi_j = O(j^{-d-1})$, it follows that $\sum_{t=P+1}^{T} y_{t-i}\tilde{v}_t = O_p\left(P^{-d}T^{2d}\right) + O_p\left(T\right)$ and $\|X'\tilde{v}\| = O_p\left(P^{-d+1/2}T^{2d}\right) + O_p\left(TP^{1/2}\right)$. Thus, overall

$$\left\|\left(X'X\right)^{-1}\right\| \|X'\tilde{v}\| = O_p\left(P^{-d+3/2}\right) + O_p\left(T^{1-2d}P^{3/2}\right) \tag{22}$$

The second term is $o_p(1)$ if, e.g., $P = O(\ln T^{\alpha})$ for all $\alpha > 0$. The first term is $o_p(P)$ since $d > 1/2$. Overall, as long as $P = o\left(T^{c}\right)$ for all $c > 0$, it follows that

$$\|\hat{\pi} - \pi\| = o_p(P) \tag{23}$$

which proves the theorem.

References

Abadir, K., Distaso, W. and L. Giraitis (2007): "Nonstationary extended local Whittle estimation," *Journal of Econometrics*, 141, 1353-1384.

Agiakoglou, C., Newbold, P. and M. Wohar (1992): "Bias in an estimator of the fractional differencing parameter," *Journal of Time Series Analysis*, 14, 235-246.

Amemiya, T. (1973): Generalized least squares with an estimated autocovariance matrix, *Econometrica*, 41, 723-732.

Andersen, T., Bollerslev, T., Diebold, F. X. and Labys (2003): Modelling and forecasting realised volatilty, *Econometrica*, 71, 579-625.

Baillie, R. T., C.-F. Chung and M. A. Tieslau (1996): "Analysing inflation by the fractionally integrated ARFIMA-GARCH model, " *Journal of Applied Econometrics*, 11, 23-40.

Baillie, R. T. and G. Kapetanios (2007): "Testing for neglected nonlinearity in long memory models, " *Journal of Business and Economic Statistics*, 25, 447-461.

Baillie, R. T., and G. Kapetanios (2008): "Nonlinear models for strongly dependent processes and financial applications, " *Journal of Econometrics*, forthcoming.

Beran, J. (1970): *Statistics for Long Memory Processes*, Chapman-Hall.

Box, G.E.P. and G.M. Jenkins (1970): *Time Series Analysis Forecasting and Control*, Holden Day.

Dahlhaus, R. (1989): "Effcient Parameter Estimation for Self-Similar Processes," *Annals of Statistics*, 17(4), 1749–1766.

Dalla, V., L. Giraitis, and J. Hidalgo (2005): "Consistent Estimation of the Long Memory Parameter for Nonlinear Time Series, " *Journal of Time Series Analysis*, 27, 211-251.

Diebold, F. X., S. Husted, and M. Rush (1991): "Real Exchange Rates Under the Gold Standard," *Journal of Political Economy*, 99, 1252–1271.

Diebold, F. X., and A. Inoue (2001): "Long Memory and Regime Switching," *Journal of Econometrics*, 105, 131–159.

Diebold, F. X. and G. Rudebusch (1989): "Long memory and persistence in aggregate output," *Journal of Monetary Economics*, 24, 189-209.

Fox, R., and M. S. Taqqu (1986): "Large Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Processes," *Annals of Statistics*, 14, 517–532.

Geweke, J., and S. Porter Hudak (1983): "The estimation and application of long memory time series models," *Journal of Time Series Analysis*, 4, 221–238.

Granger, C. W. J. (1980): "Long Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227–238.

Granger, C. W. J., and R. Joyeux (1980): "An Introduction to Long Memory Time Series Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 15–39.

Hamilton, J. (1994): *Time Series Analysis*, Princeton University Press.

Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society (Series B)*, 41, 190–195.

Henry, M. (2001): "Robust automatic bandwidth for long memory," *Journal of Time Series Analysis*, 22, 292-316.

Henry, M., and P. M. Robinson (1996): "Bandwidth Choice in Gaussian Semiparametric Estimation of Long Range Dependence," in *Athens Conference on Applied Probability and Time Series, Volume II: Time Series in Memory of E. J. Hannan*, ed. by P. M. Robinson, and M. Rosenblatt. Springer-Verlag.

Hosking, J. R. M. (1981): "Fractional Differencing," *Biometrika*, 65, 165–176.

Hosoya, Y. (1997): "A Limit Theory for Long Range Dependence and Statistical Inference on Related Models," *Annals of Statistics*, 25, 105–137.

Hurst, H. E. (1951): "Long Term Storage Capacity of Reservoirs," *Transactions of the American Society of Civil Engineers*, 116, 770–799.

Inoue, A. and Y. Kasahara (2006): "Explicit representation of finite predictor coefficients and its applications," *Annals of Statistics*, 34, 973–993.

Jensen, M. (2007): "The long-run Fisher effect: can it be tested?," *Journal of Money, Credit and Banking*, forthcoming.

Johansen, S. and M. O. Nielsen (2008): "Likelihood Inference for a Nonstationary Fractional Autoregressive Model" *Mimeo*.

Li, W.-K. and W. K. McLeod (1986): Fractional time series modeling *Biometrika*, 73, 217-221.

Ling, S. and W. K. Li (2001): "Asymptotic Inference for Nonstationary Fractionally Integrated Autoregressive Moving-Average Models" *Econometric Theory*, 17, 738-765.

Lo, A. W. (1991): Long term memory in stock market prices *Econometrica*, 59, 1279-1313.

Moehring, R. (1990): Parameter estimation in gaussian intermediate memory time series *University of Hamburg working paper*.

Nielsen, M. O. and P. H. Frederiksen (2004): "Finite Sample Comparison of Parametric, Semiparametric, and Wavelet Estimators of Fractional Integration" *Econometric Reviews*, 24, 405-443.

Phillips, P.C.B. (2007): "Unit Root Periodogram," *Journal of Econometrics,* 138, 104-124.

Phillips, P.C.B. and K. Shimotsu (2006): "Local Whittle estimation in non stationary and unit root cases," *Annals of Statistics,* 32, 656-692.

Phillips, P.C.B. and K. Shimotsu (2006): "Local Whittle Estimation of Fractional Integration and Some of its Variants," *Journal of Econometrics,* 130, 209-233.

Poskitt, D. (2005): "Autoregressive Approximation in Nonstandard Situations: The Non-Invertible and Fractionally Integrated Case," *Annals of the Institute of Statistical Mathematics,* 59, 697–725.

Robinson, P. M. (1995): "Gaussian Semi-parametric Estimation of Long Range Dependence," *Annals of Statistics*, 23, 1630–1661.

Robinson, P. M. (2006): "Conditional - Sum - of - Squares Estimation of Models for Stationary Time Series with Long Memory," *London School of Economics working paper.*

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *Annals of Statistics*, 8, 147–164.

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

Sowell, F. B. (1992): "Modeling long run behavior with the fractional ARIMA model" *Journal of Monetary Economics*, 29, 277-302.

Tanaka, K. (1999): "The Nonstationary Fractional Unit Root" *Econometric Theory*, 15, 549–583.

Velasco, C. (1999): "Gaussian semi parametric estimation of non stationary time series," *Journal of Time Series Analysis*, 20, 87-127.

Wright, J H (1995): "Stochastic orders of magnitude associated with two-stage estimators of fractional ARMA systems," *Journal of Time Series Analysis*, 16, 119-125

Table 1. Results for estimated d												
Bias												
	T=200				T=400				T=1000			
d/AR	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99
	Two-Step (m=0.5)											
0	0.005	0.477	0.700	0.947	-0.002	0.333	0.592	0.919	-0.001	0.254	0.495	0.887
0.1	0.016	0.476	0.713	0.936	0.001	0.338	0.603	0.899	0.000	0.258	0.499	0.865
0.2	0.030	0.474	0.687	0.882	0.010	0.351	0.586	0.867	0.005	0.252	0.499	0.852
0.3	0.036	0.479	0.692	0.818	0.012	0.343	0.586	0.822	0.008	0.258	0.491	0.829
0.4	0.036	0.480	0.660	0.744	0.024	0.350	0.579	0.756	0.007	0.255	0.492	0.766
0.49	0.076	0.469	0.622	0.662	0.042	0.352	0.557	0.681	0.017	0.257	0.483	0.719
	Two-Step (optimal m)											
0	0.041	0.136	0.168	0.578	0.021	0.033	0.140	0.451	0.031	0.063	0.064	0.173
0.1	0.034	0.096	0.093	0.582	0.018	0.057	0.127	0.418	0.026	0.057	0.090	0.270
0.2	0.040	0.181	0.163	0.515	0.028	0.050	0.129	0.396	0.032	0.062	0.080	0.196
0.3	0.053	0.114	0.045	0.520	0.030	0.087	0.175	0.341	0.031	0.066	0.087	0.163
0.4	0.066	0.190	0.116	0.457	0.045	0.088	0.195	0.333	0.037	0.062	0.098	0.139
0.49	0.080	0.211	0.082	0.396	0.051	0.107	0.190	0.310	0.039	0.077	0.083	0.098
	Maximum Likelihood											
0	-0.057	0.025	0.027	0.016	-0.038	0.019	0.007	0.007	-0.023	0.007	0.004	0.003
0.1	-0.059	0.033	0.009	-0.014	-0.042	0.016	0.008	-0.012	-0.017	0.006	0.002	-0.005
0.2	-0.040	0.017	0.005	-0.060	-0.037	0.016	-0.003	-0.052	-0.014	0.002	-0.006	-0.028
0.3	-0.047	0.007	-0.048	-0.044	-0.036	-0.002	-0.033	-0.085	-0.018	-0.003	-0.017	-0.077
0.4	-0.041	-0.050	-0.102	0.039	-0.025	-0.033	-0.083	-0.068	-0.017	-0.018	-0.047	-0.124
0.49	-0.029	-0.112	-0.154	0.078	-0.023	-0.079	-0.138	0.010	-0.011	-0.044	-0.092	-0.121
Variance												
	T=200				T=400				T=1000			
d/AR	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99
	Two-Step (m=0.5)											
0	0.206	0.218	0.209	0.183	0.166	0.170	0.173	0.155	0.115	0.113	0.117	0.115
0.1	0.212	0.207	0.205	0.170	0.164	0.171	0.176	0.145	0.115	0.113	0.121	0.117
0.2	0.203	0.217	0.205	0.152	0.168	0.171	0.169	0.141	0.113	0.114	0.118	0.122
0.3	0.212	0.208	0.199	0.146	0.165	0.178	0.167	0.131	0.114	0.117	0.122	0.137
0.4	0.205	0.201	0.185	0.135	0.163	0.165	0.158	0.134	0.110	0.116	0.119	0.159
0.49	0.192	0.192	0.179	0.133	0.157	0.159	0.161	0.132	0.114	0.117	0.126	0.177
	Two-Step (optimal m)											
0	0.183	0.574	1.072	1.045	0.133	0.400	0.567	1.061	0.072	0.212	0.343	1.177
0.1	0.184	0.598	1.173	1.055	0.131	0.402	0.590	1.060	0.070	0.205	0.356	1.098
0.2	0.176	0.548	1.076	1.020	0.125	0.391	0.563	1.043	0.074	0.208	0.357	1.074
0.3	0.172	0.601	1.112	0.961	0.125	0.397	0.549	1.091	0.077	0.211	0.349	1.107
0.4	0.187	0.545	1.070	0.945	0.127	0.405	0.579	1.035	0.071	0.220	0.336	1.112
0.49	0.171	0.554	1.132	0.893	0.128	0.385	0.574	0.988	0.075	0.209	0.341	1.043
	Maximum Likelihood											
0	0.181	0.142	0.122	0.076	0.142	0.106	0.077	0.049	0.088	0.061	0.042	0.030
0.1	0.178	0.145	0.116	0.080	0.143	0.104	0.080	0.048	0.083	0.062	0.043	0.031
0.2	0.177	0.144	0.123	0.116	0.141	0.110	0.069	0.061	0.087	0.059	0.040	0.050
0.3	0.174	0.150	0.128	0.271	0.137	0.102	0.079	0.177	0.088	0.058	0.043	0.070
0.4	0.181	0.144	0.160	0.395	0.139	0.109	0.103	0.305	0.085	0.060	0.061	0.179
0.49	0.182	0.172	0.217	0.430	0.137	0.132	0.158	0.395	0.091	0.075	0.093	0.282

Table 2. Results for estimated AR coefficient												
Bias												
	T=200				T=400				T=1000			
d/AR	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99
	Two-Step ($m = 0.5$)											
0	-0.030	-0.447	-0.662	-0.924	-0.015	-0.286	-0.540	-0.897	-0.003	-0.194	-0.411	-0.866
0.1	-0.041	-0.440	-0.683	-0.923	-0.016	-0.294	-0.545	-0.880	-0.005	-0.200	-0.417	-0.841
0.2	-0.053	-0.443	-0.652	-0.871	-0.021	-0.302	-0.527	-0.850	-0.010	-0.195	-0.419	-0.826
0.3	-0.054	-0.447	-0.662	-0.796	-0.024	-0.299	-0.529	-0.795	-0.011	-0.200	-0.408	-0.794
0.4	-0.053	-0.445	-0.618	-0.705	-0.035	-0.304	-0.520	-0.707	-0.011	-0.199	-0.408	-0.711
0.49	-0.088	-0.433	-0.575	-0.595	-0.053	-0.303	-0.493	-0.605	-0.021	-0.199	-0.400	-0.650
	Two-Step (optimal m)											
0	-0.061	-0.247	-0.415	-0.657	-0.032	-0.126	-0.240	-0.579	-0.035	-0.075	-0.114	-0.458
0.1	-0.055	-0.231	-0.414	-0.667	-0.029	-0.138	-0.245	-0.569	-0.031	-0.069	-0.130	-0.482
0.2	-0.058	-0.272	-0.425	-0.636	-0.036	-0.131	-0.237	-0.551	-0.037	-0.074	-0.125	-0.438
0.3	-0.073	-0.243	-0.391	-0.630	-0.041	-0.155	-0.262	-0.546	-0.035	-0.077	-0.127	-0.443
0.4	-0.082	-0.280	-0.409	-0.594	-0.054	-0.159	-0.285	-0.531	-0.042	-0.077	-0.130	-0.438
0.49	-0.092	-0.299	-0.418	-0.535	-0.060	-0.166	-0.285	-0.510	-0.044	-0.085	-0.124	-0.397
	Maximum Likelihood											
0	0.035	-0.034	-0.029	-0.013	0.023	-0.023	-0.011	-0.006	0.018	-0.007	-0.005	-0.002
0.1	0.035	-0.040	-0.023	-0.012	0.028	-0.022	-0.011	-0.005	0.014	-0.008	-0.004	-0.002
0.2	0.018	-0.033	-0.027	-0.016	0.025	-0.021	-0.008	-0.003	0.010	-0.006	-0.002	-0.002
0.3	0.028	-0.035	-0.013	-0.102	0.025	-0.014	-0.003	-0.034	0.014	-0.003	-0.001	0.001
0.4	0.024	-0.009	-0.012	-0.287	0.014	-0.004	0.004	-0.134	0.014	0.000	0.005	-0.030
0.49	0.012	0.005	-0.027	-0.419	0.010	0.008	-0.000	-0.306	0.007	0.008	0.010	-0.107
Variance												
	T=200				T=400				T=1000			
d/AR	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99	0.5	0.9	0.95	0.99
	Two-Step ($m = 0.5$)											
0	0.189	0.223	0.230	0.198	0.157	0.165	0.201	0.171	0.116	0.105	0.138	0.135
0.1	0.195	0.213	0.222	0.184	0.157	0.169	0.205	0.167	0.116	0.105	0.143	0.141
0.2	0.190	0.222	0.228	0.174	0.163	0.174	0.199	0.168	0.114	0.105	0.141	0.150
0.3	0.196	0.214	0.220	0.172	0.161	0.176	0.192	0.160	0.115	0.108	0.144	0.170
0.4	0.194	0.211	0.210	0.169	0.157	0.165	0.186	0.170	0.111	0.108	0.143	0.209
0.49	0.184	0.203	0.203	0.173	0.159	0.163	0.190	0.172	0.115	0.108	0.149	0.238
	Two-Step (optimal m)											
0	0.175	0.356	0.539	0.585	0.132	0.247	0.355	0.579	0.074	0.134	0.195	0.565
0.1	0.175	0.351	0.540	0.587	0.129	0.248	0.355	0.571	0.075	0.124	0.218	0.569
0.2	0.167	0.359	0.527	0.571	0.127	0.243	0.349	0.566	0.078	0.130	0.215	0.541
0.3	0.173	0.352	0.510	0.551	0.126	0.249	0.345	0.557	0.080	0.131	0.216	0.537
0.4	0.179	0.354	0.510	0.528	0.129	0.259	0.361	0.542	0.075	0.137	0.210	0.533
0.49	0.165	0.365	0.514	0.513	0.129	0.246	0.356	0.528	0.081	0.130	0.202	0.508
	Maximum Likelihood											
0	0.183	0.107	0.086	0.040	0.142	0.080	0.048	0.020	0.092	0.041	0.019	0.006
0.1	0.179	0.110	0.082	0.045	0.145	0.077	0.049	0.012	0.089	0.043	0.023	0.006
0.2	0.180	0.109	0.084	0.096	0.146	0.081	0.037	0.011	0.091	0.038	0.016	0.032
0.3	0.177	0.114	0.083	0.292	0.141	0.073	0.039	0.178	0.094	0.035	0.016	0.005
0.4	0.183	0.092	0.123	0.451	0.141	0.069	0.045	0.343	0.089	0.035	0.017	0.180
0.49	0.185	0.095	0.187	0.502	0.143	0.072	0.116	0.471	0.095	0.033	0.036	0.322

Table 3. Optimal Bandwidths for $AR(1)$ with α reported where $m = T^{\alpha}$					
d/ϕ	0.50	0.70	0.80	0.90	0.95	0.99
0.00	0.59	0.52	0.46	0.37	0.29	0.10
0.10	0.59	0.52	0.46	0.37	0.29	0.10
0.20	0.59	0.52	0.46	0.37	0.29	0.10
0.30	0.59	0.52	0.46	0.37	0.29	0.10
0.40	0.59	0.52	0.46	0.37	0.29	0.10
0.49	0.60	0.52	0.46	0.37	0.29	0.10

Table 4. Results for $ARFIMA(2,d,0)$.						
Parameter/d	0.000	0.100	0.200	0.300	0.400	0.490
Bias						
LW						
d	0.617	0.613	0.610	0.609	0.608	0.594
ϕ_1	0.001	0.021	0.046	0.088	0.158	0.274
ϕ_2	-0.001	-0.010	-0.031	-0.071	-0.141	-0.257
MLE						
d	0.002	0.003	0.004	0.005	-0.002	-0.023
ϕ_1	0.000	0.000	0.001	0.005	0.015	0.040
ϕ_2	-0.001	-0.002	-0.003	-0.008	-0.018	-0.042
Variances						
LW						
d	0.011438	0.008869	0.004104	0.012011	0.008045	0.008189
ϕ_1	0.000014	0.000033	0.000283	0.000641	0.000882	0.004585
ϕ_2	0.000016	0.000021	0.000299	0.000641	0.000901	0.004612
MLE						
d	0.000146	0.000175	0.000269	0.000208	0.000183	0.000637
ϕ_1	0.000014	0.000054	0.000079	0.000143	0.000305	0.001890
ϕ_2	0.000014	0.000015	0.000173	0.000212	0.000356	0.002153

Figure 1: Impulse Responses: d=0.3, ar=0.5

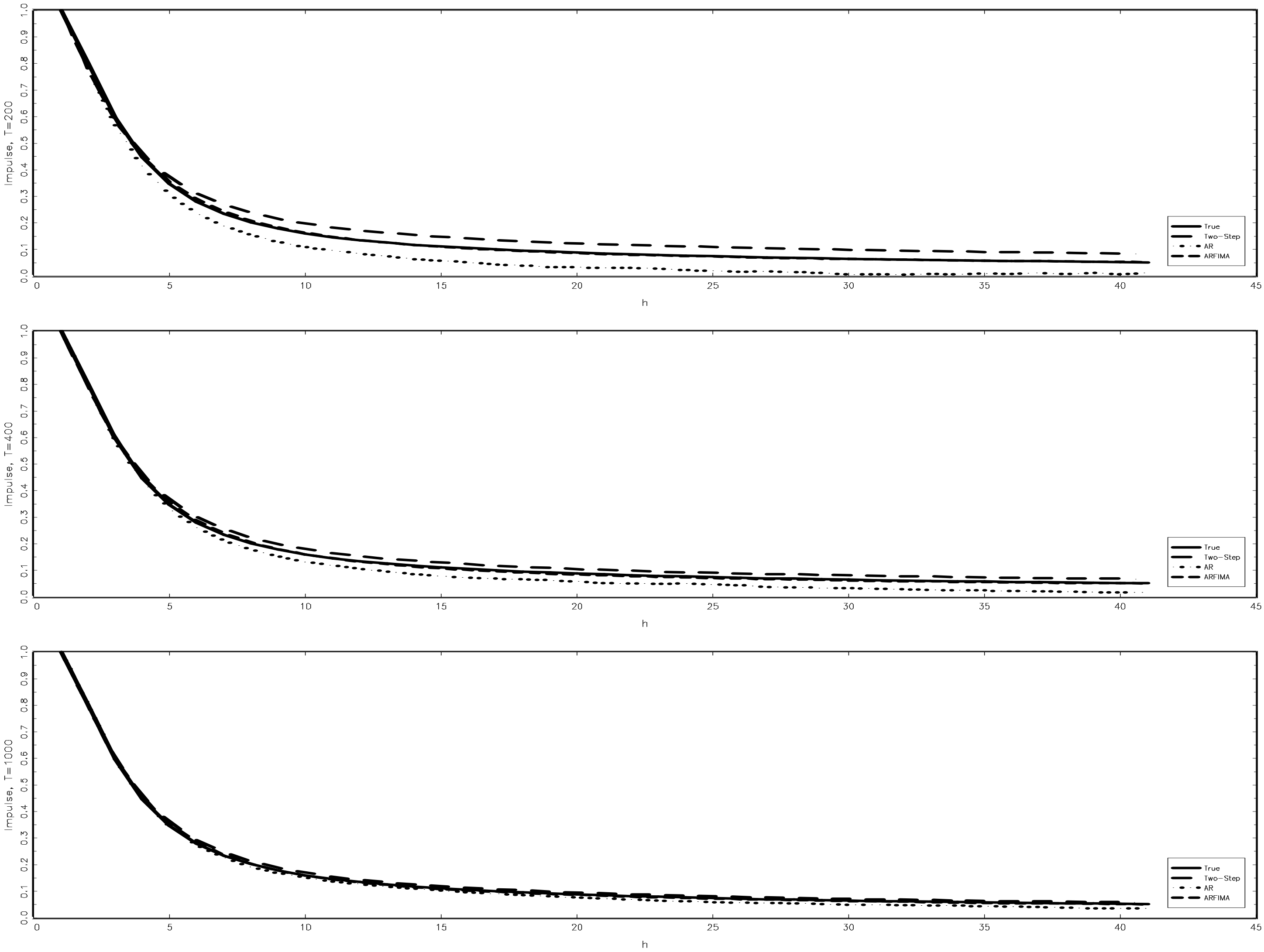

Figure 2: Impulse Responses: d=0.3, ar=0.95

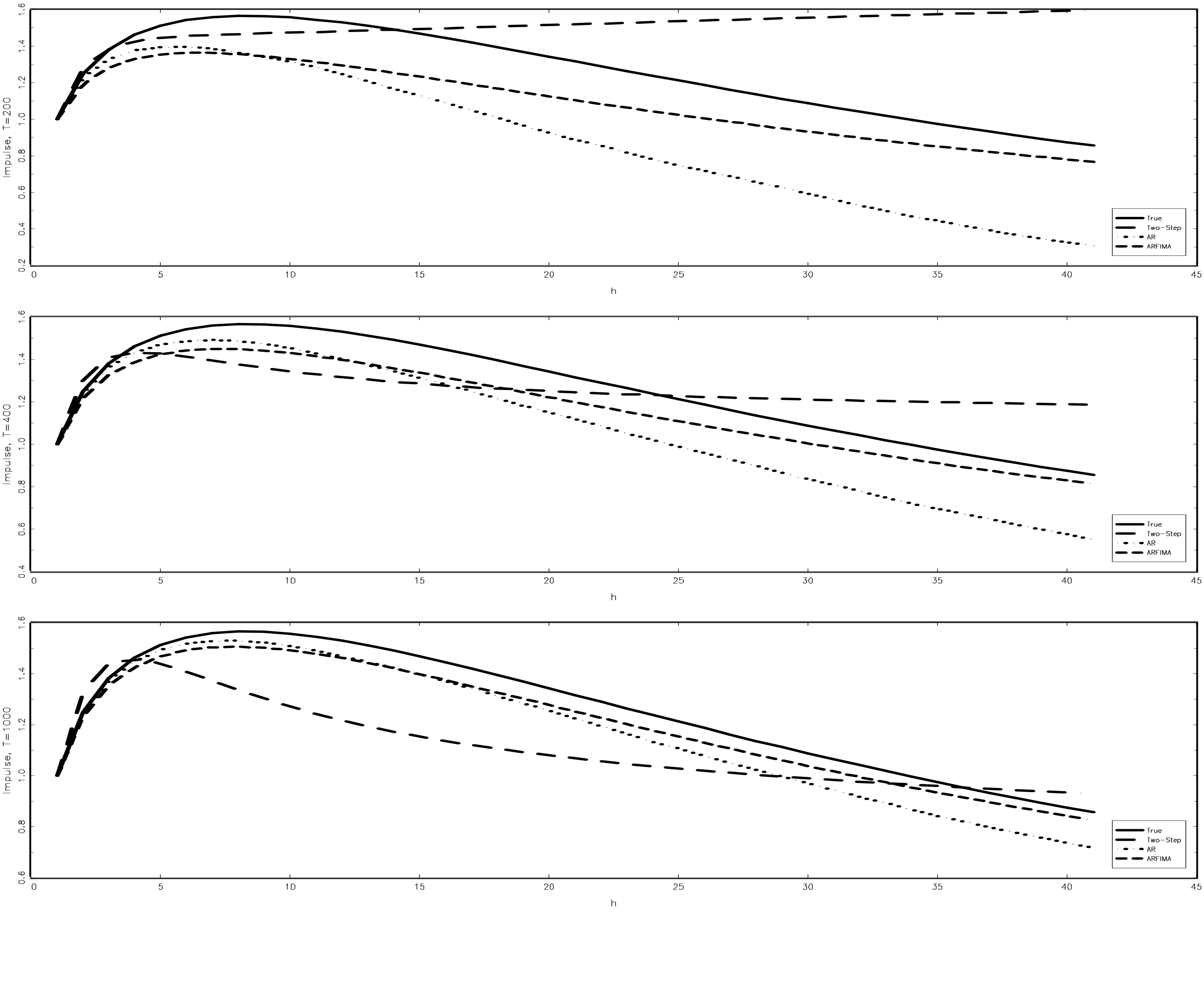

Figure 3: Impulse Responses: d=0.49, ar=0.5

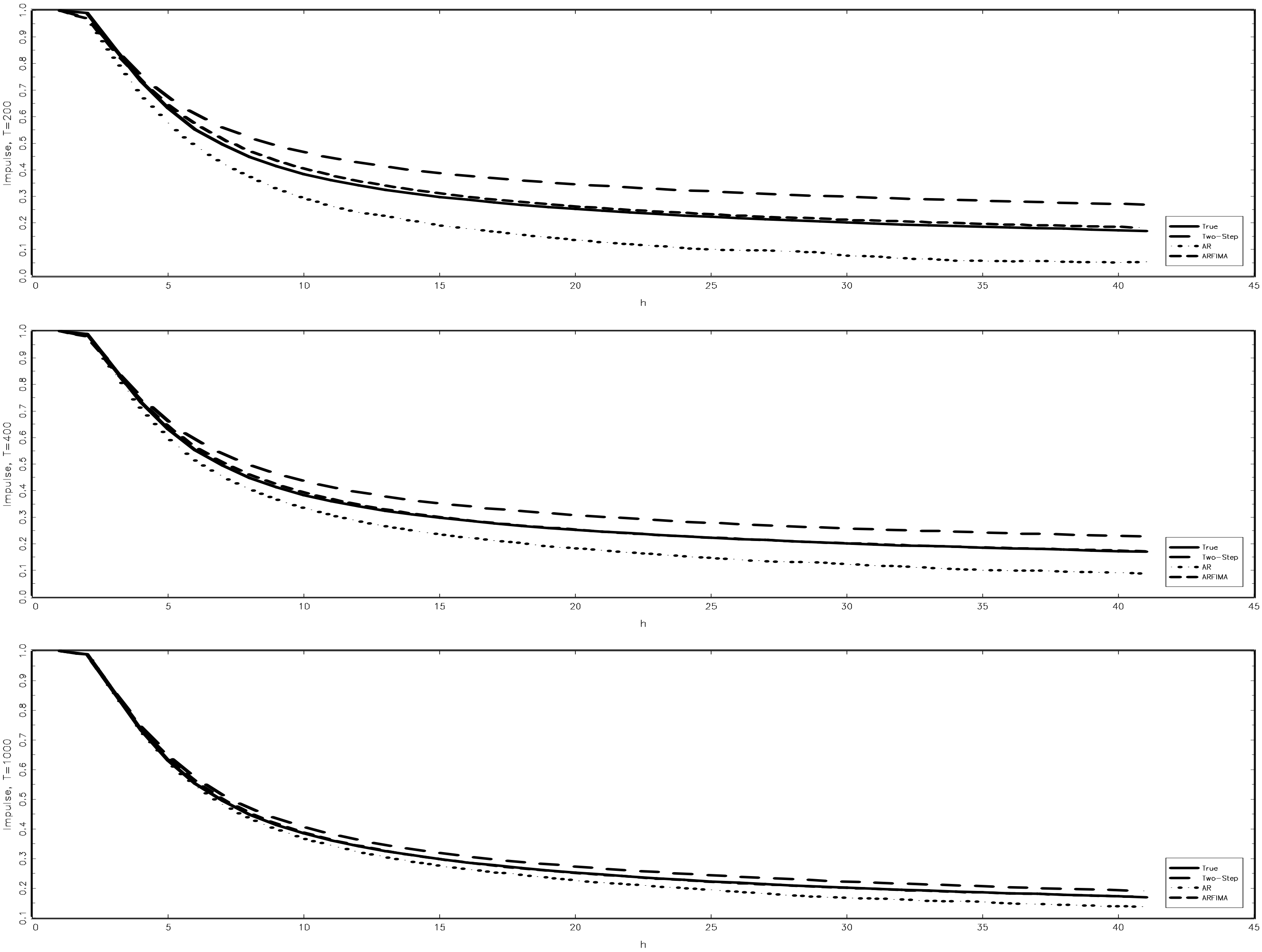

Figure 4: Impulse Responses: d=0.49, ar=0.95

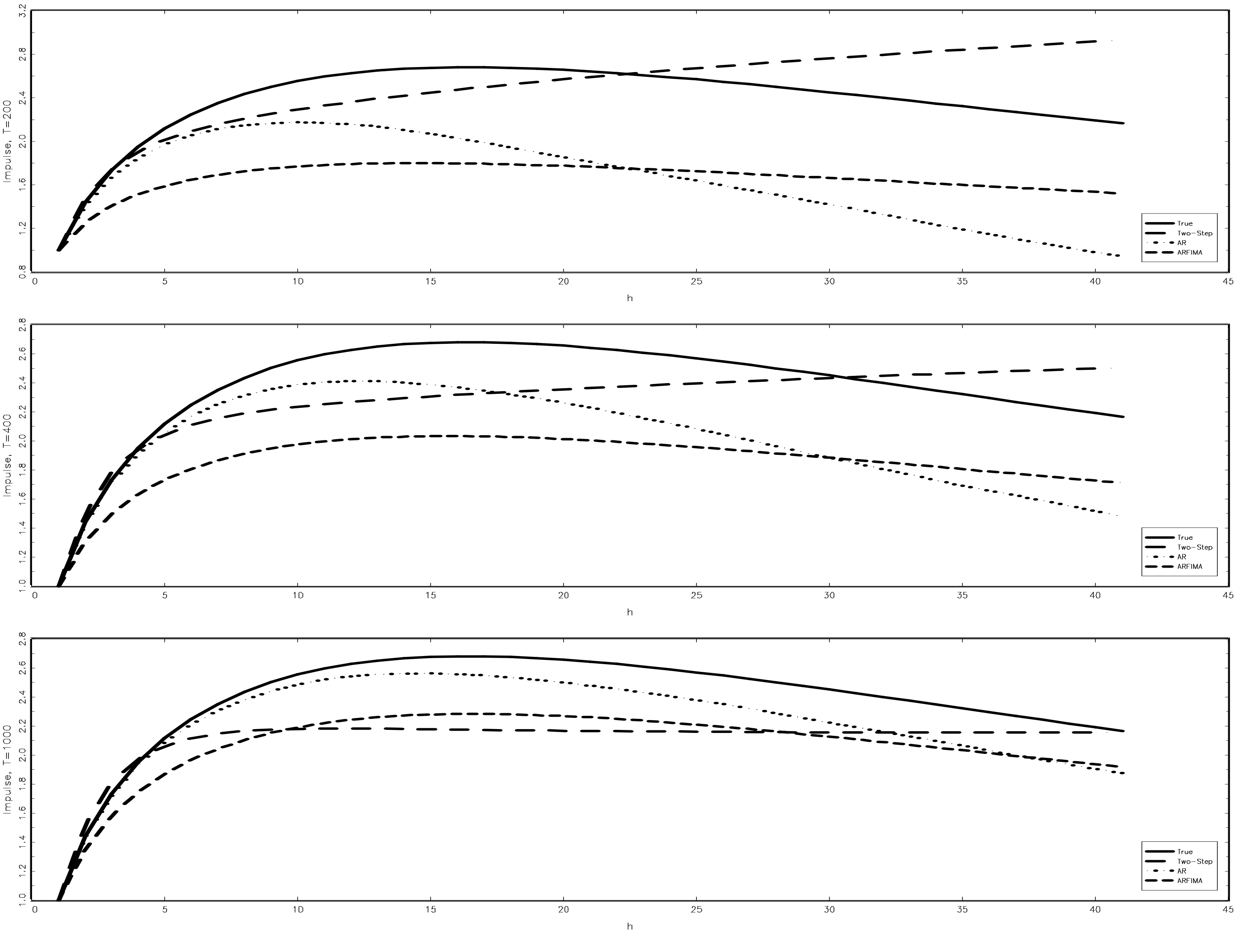

Figure 5: Impulse Responses: d=0.6, ar=0.95

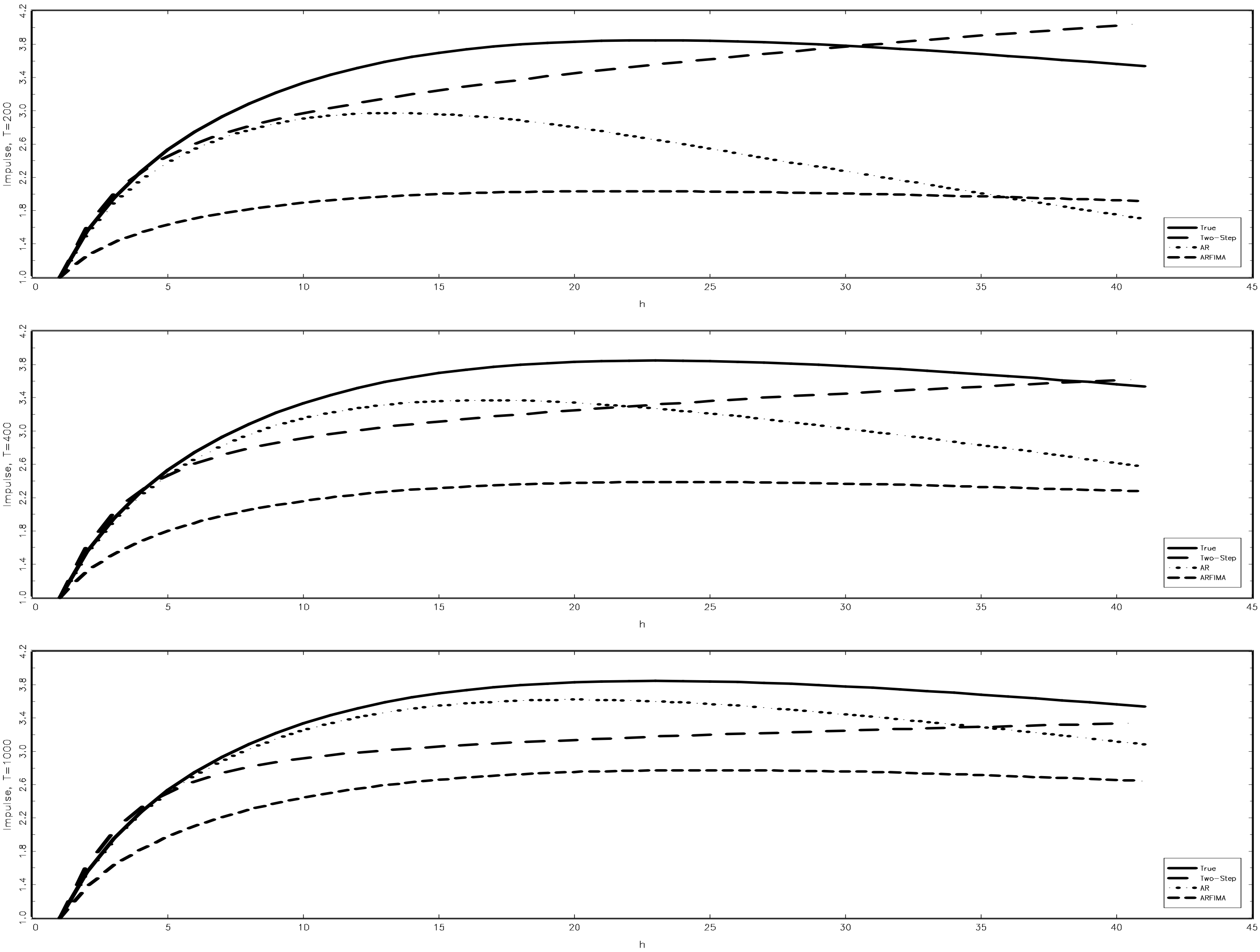

Figure 6: Impulse Responses: d=0.8, ar=0.5

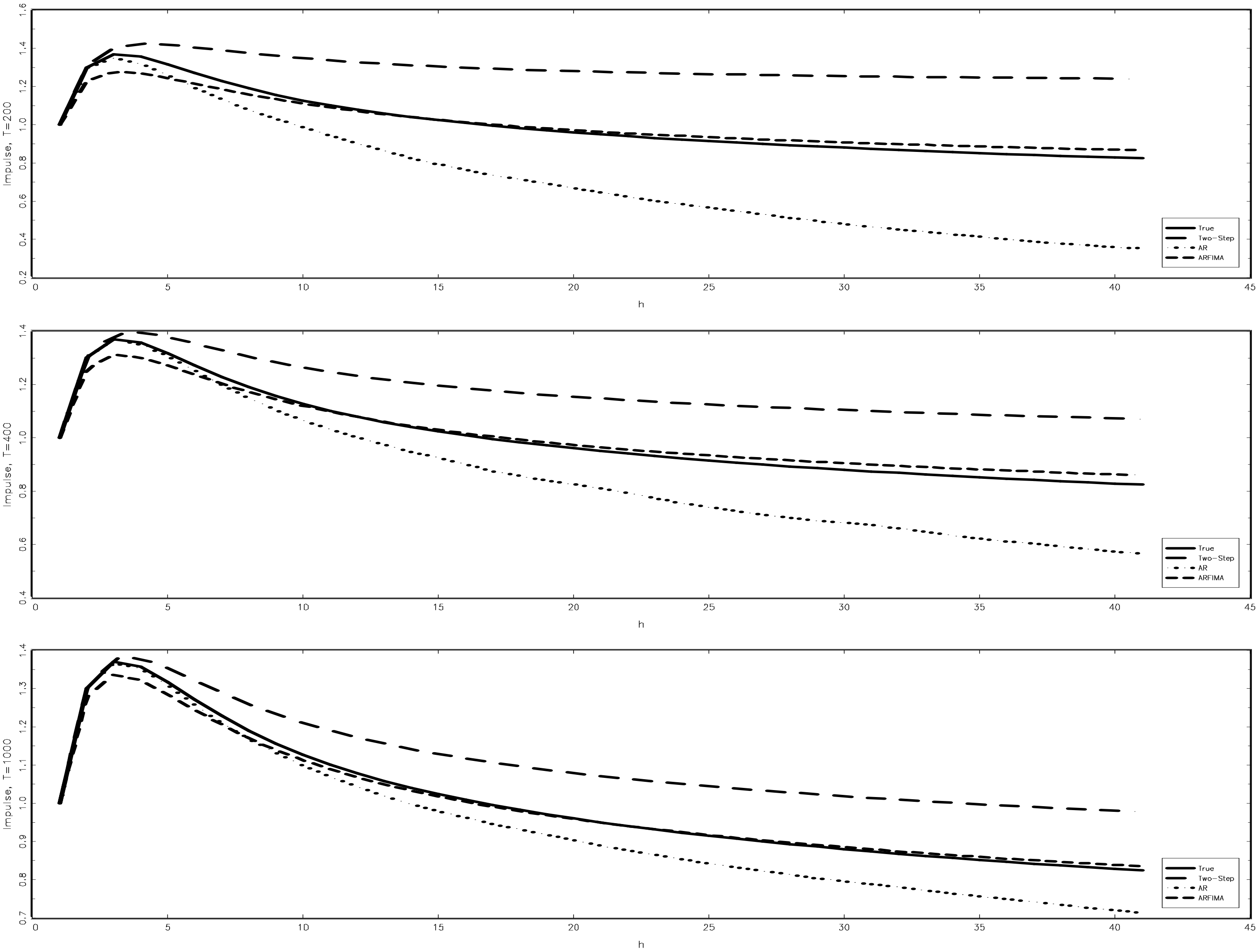

Figure 7: Impulse Responses: d=0.8, ar=0.95