PROPERTIES OF PREDICTORS IN OVERDIFFERENCED NEARLY NONSTATIONARY AUTOREGRESSION*

Ismael Sánchez and Daniel Peña

WP-AD 99-08

Correspondence to I. Sánchez: Universidad de Alicante. Departamento de Fundamentos del Análisis Económico. Ap. Correos 99, 03080 Alicante.

Editor: Instituto Valenciano de Investigaciones Económicas, s.a.
Primera Edición Mayo 1999
ISBN: 84-482-2121-4
Depósito Legal: V-2104-1999

* I. Sánchez: Universidad de Alicante, D. Peña: Universidad Carlos III de Madrid.

PROPERTIES OF PREDICTORS IN OVERDIFFERENCED NEARLY NONSTATIONARY AUTOREGRESSION

Ismael Sánchez and Daniel Peña

A B S T R A C T

This paper analyzes the effect of overdifferencing a stationary AR(p+1) process whose largest root is near unity. It is found that if the process is nearly nonstationary, the estimators of the overdifferenced model ARIMA (p, 1, 0) are root-T consistent. It is also found that this misspecified ARIMA (p, 1, 0) has lower predictive mean squared error, to terms of small order, that the properly specified AR(p+1) model due to its parsimony. The advantage of the overdifferenced predictor depends on the remaining roots, the prediction horizon, and the mean of the process.

Keywords: Autoregressive processes, near nonstationarity, overdifferencing, parsimony, predictive mean squared error, unit roots.

1 Introduction

In this paper, we investigate the consequences in estimation and prediction of overdifferencing a stationary AR(p+1) with a root close to unity. Differencing is normally used to transform a homogeneous linear nonstationary time series into a stationary process that is often modeled as an ARMA(p, q) process. It is said, then, that the original series follows an ARIMA(p, d, q) process, where d is the number of differences required to obtain stationarity. We assume that the process is not a long memory process (see, for instance, Granger & Joyeux, 1980) and, thus, d is an integer equal to the number of unit roots in the autoregressive characteristic equation. When a stationary process has an autoregressive characteristic equation with a root close to unity it is said to be nearly nonstationary. Given a small or moderate sample of this process, it is very likely to conclude, due to the low power of unit roots tests in this case, that a difference should be applied. The differenced series will be noninvertible and the process is called overdifferenced.

Since the work of Fuller (1976) and Dickey & Fuller (1979), there has been a vast literature concerning the detection of unit roots in autoregressive polynomials. This literature notes the difficulty of a correct detection in near nonstationary processes. In spite of this, relatively little has been written on the consequences of a wrong detection. Previous work on the effect of overdifferencing can be found in Plosser & Schwert (1977, 1978), Harvey (1981), Campbell & Perron (1991), and Stock (1996). Plosser & Schwert (1977) examine, using Monte Carlo techniques, the effect of overdifferencing in two cases: processes with a deterministic linear trend and stochastic regression models. They conclude that, in these situations, the loss in efficiency on both parameter estimation and prediction is not substantial, provided an MA parameter is included. Harvey (1981), assuming known parameters, also concludes that overdifferencing does not need to have serious implications for prediction, provided a finite sample prediction procedure is used and an MA parameter is included. Campbell & Perron (1991) and Stock (1996) compare, using simulations by Monte Carlo, the prediction accuracy of an AR(1) and a random walk. The empirical results of these authors show that the random walk can produce forecasts with lower prediction mean squared error (PMSE) than the AR(1) if the root is close to unity.

In this paper, we justify theoretically the advantages of the overdifferenced predictor, found empirically by Campbell & Perron (1991) and Stock (1996), in a general autoregression and analyze the effect of other factors like the remaining roots, sample size (T), and horizon (H). We will assume that a root of the AR($p+1$) is close to unity and, thus, we will adopt as a more plausible overdifferenced predictor the ARIMA($p, 1, 0$) model, where no MA component is involved.

We will prove that the PMSE of the overdifferenced model ARIMA($p, 1, 0$) is lower, to terms of small order, than the PMSE of the correct model AR($p + 1$) if the root that is closer to unity, ρ^{-1}, follows $\rho = \exp(-c/T^{\beta})$; $\beta > 1$. The advantage of the overdifferenced predictor is due to its parsimony. Therefore, it is larger if the AR(p+1) process has a non-zero mean, since it will vanish in the overdifferenced model. The remaining roots also affect the advantage of the overdifferenced predictor. Positive roots increase the advantage of the overdifferenced model, whereas negative roots have the opposite effect. The advantage of the overdifferenced model is small in the short term, but can increase with the horizon.

An important consequence of these results is that, for forecasting purposes, it is better to overdifferentiate than to underdifferentiate. Therefore, the possible low power of unit root tests in autoregression is not as important in forecasting as in model identification, since we can still obtain an efficient predictor.

This paper is organized as follows. In Section 2 we introduce the model and notation. In Section 3 we define nearly nonstationary processes. The consequences of overdifferencing in estimation are analyzed in section 4, and the effect on the PMSE for each predictor in section 5. In Section 6 we compare the PMSE of the competing models and extract further results from the AR(1) case. A simulation study is presented in section 7 to illustrate the results.

2 The model and notation

Let $\{y_t\}$ be the following stationary AR$(p+1)$ process:

$$\varphi(B)y_t = \phi(B)(1-\rho B)y_t = \alpha + a_t, \tag{2.1}$$

where B is the backshift operator; $\varphi(B) = (1 - \sum_{i=1}^{p+1} \varphi_i B^i)$ is a polynomial operator on B such that $\varphi(B) = 0$ has all its roots outside the unit circle, with ρ^{-1} being the closer to unity root. Let a_t be a sequence of independent identically distributed random variables with zero mean and variance σ^2. Let $\mu = \mathrm{E}(y_t)$; then $\alpha = \mu\varphi(1)$. We make the following assumption:

A1. For some $s_0 > 2$, $E\{|a_t|^{s_0}\} < \infty$.

It is well known that this model can be represented in first-order vector autoregressive form as follows:

$$Y_t = A_\alpha Y_{t-1} + U_{t,p+2}, \tag{2.2}$$

with $Y_t = (y_t, ..., y_{t-p}, 1)'$, $U_{t,p+2} = (a_t, 0, ..., 0)'$, where the subindex $(p+2)$ indicates the dimension of the vector and

$$A_\alpha = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_p & \varphi_{p+1} & \alpha \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

Then $y_t = e'_{p+2} Y_t$, with $e_{p+2} = (1, 0, ..., 0)'$. Let $\Gamma_y = E(Y_t Y'_t)$ and $\gamma_y = E(Y_t y_{t+1})$. If we represent the process in deviations from the mean, we obtain $\tilde{Y}_t = A_o \tilde{Y}_{t-1} + U_{t,p+1}$, where $\tilde{Y}_t = (\tilde{y}_t, \tilde{y}_{t-1}, ..., \tilde{y}_{t-p})'$, $\tilde{y}_t = y_t - \mu$, and A_o is the first $(p+1) \times (p+1)$ submatrix of A_α. We will also denote $\Gamma_{\tilde{y}} = E(\tilde{Y}_t \tilde{Y}'_t)$. If a difference is applied to y_t, the series obtained, $w_t = (1 - B) y_t$, can be represented as

$$\phi(B)(1 - \rho B) w_t = (1 - B) a_t, \tag{2.3}$$

which is noninvertible. The process w_t has the following vector representation (Lütkepohl, 1991, p. 223)

$$Z_t = A_1 Z_{t-1} + U^*_{t,p+2}, \tag{2.4}$$

with $Z_t = (W'_t, a_t)'$, $W_t = (w_t, ..., w_{t-p})'$, $U^*_{t,p+2} = (a_t, 0, \ldots, 0, a_t)'$, and

$$A_1 = \begin{pmatrix} A_o & -e_{p+1} \\ 0 \cdots 0 & 0 \end{pmatrix}$$

with $w_t = e'_{p+1} Z_t$. Let $\Gamma_w = E(W_t W'_t)$ and $\gamma_w = E(W_t w_{t+1})$. In what follows, we will use the hat symbol ($\hat{o}$) to denote estimates from a sample of the overdifferenced process $\{w_t\}$ and the check symbol ($\check{o}$) for estimates from a sample of the original process $\{y_t\}$. The least squares estimator of the AR(p+1) parameter vector $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_{p+1}, \alpha)'$, fitted to a sample of size T of the original process, is $\check{\boldsymbol{\varphi}} = \check{\Gamma}_y^{-1} \check{\gamma}_y$, where $\check{\Gamma}_y = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} Y_j Y'_j$ and $\check{\gamma}_y = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} Y_j y_{j+1}$. Similarly, the least squares estimator of the parameter vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)'$ of a misspecified AR(p), fitted to a sample of size $T-1$ of the overdifferenced process (2.3), is $\hat{\boldsymbol{\phi}} = \hat{\Gamma}_w^{-1} \hat{\gamma}_w$, where $\hat{\Gamma}_w = (T-p-1)^{-1} \sum_{j=p+1}^{T-1} W_j W'_j$ and $\hat{\gamma}_w = (T - p - 1)^{-1} \sum_{j=p+1}^{T-1} W_j w_{j+1}$. We also make the following assumptions, where $\|\cdot\|$ denotes the Euclidean norm:

A2. $E(\|\check{\Gamma}_y^{-1}\|^{2k})$ $(k = 1, 2, ..., k_0)$ is bounded for all finite and sufficiently large T and some k_0.
A3. $E(\|\hat{\Gamma}_w^{-1}\|^{2k})$ $(k = 1, 2, ..., k_0)$ is bounded for all finite and sufficiently large T and some k_0.

Assumptions A2 and A3 are similar to assumption A3 of Kunitomo & Yamamoto (1985). They are also equivalent to assumption A3 of Bhansali (1981). It should be noted that they are satisfied if the distribution is normal (see Fuller & Hasza, 1981). These assumptions are needed in several parts of this work, especially in application to the results of Kunitomo & Yamamoto (1985) and Bhansali (1981). They imply that, for a large enough sample size, the estimation of the covariance matrices are sufficiently near the true values (Bhansali, 1981, p. 590).

3 Nearly nonstationary autoregressions

A process is said to be nearly nonstationary (near integrated) if its autoregressive characteristic equation has a root, ρ^{-1}, very close to unity. If ρ is close enough to unity, the term $(1-\rho B)$ in (2.3) will be similar to $(1-B)$. Therefore, although the overdifferenced process w_t is strictly a noninvertible ARMA$(p+1,1)$, an average correlogram of w_t will suggest estimating by an AR(p) instead.

The similarity between w_t and a true AR(p) process does not only depend on ρ but it is influenced by the remaining roots. In order to see this point, let π_j be the coefficients of the polynomial $\pi(B) = (1-\pi_1 B - \pi_2 B^2 - \cdots)$, where $\varphi(B) = \pi(B)(1-B)$. These coefficients follow

$$\pi_j = \begin{cases} \phi_j + (\rho-1)(1-\sum_{k=1}^{j-1}\phi_k) & \text{if } j \le p, \\ (\rho-1)(1-\sum_{k=1}^{p}\phi_k) & \text{if } j > p, \end{cases} \tag{3.1}$$

with $\phi_k = 0$ if $k < 1$. If we denote as r_i^{-1}, $i = 1,\ldots,p$, to the roots of the characteristic equation $\phi(B) = 0$, then

$$\left(1 - \sum_{k=1}^{p}\phi_k\right) = \prod_{i=1}^{p}(1-r_i). \tag{3.2}$$

Therefore, negative values of r_i increase the value of π_j, $j > p$, and decrease the similarity of w_t and an AR(p).

Thus, the definition of a nearly nonstationary process needs, (1) a parameterization that converges to the unit root with the sample size and (2) a constant term that can reflect the influence of the remaining roots in finite samples. Phillips (1987) and Chan & Wei (1987) define nearly nonstationary process for the AR(1) case by reparameterizing $\rho = \exp(-c/T) = 1 - c/T + o(T^{-1})$, where c is a fixed constant. In this definition, the convergence rate to unity is fixed to be $O(T^{-1})$. These authors use this definition to provide asymptotic theory for the estimation of ρ. The formulation is justified by Phillips (1987) because this is the order of consistency of the least squares estimator, and by Chan & Wei (1987) because this is the order of the observed Fisher information of ρ under normality. In order to analyze the consequences of overdifferencing with different convergence rates we will define ρ as

$$\rho = \exp\left(-\frac{c}{T^{\beta}}\right), \tag{3.3}$$

with c and β being fixed constants. We deal only with stationary processes, and hence $c, \beta > 0$. Time series generated by (2.1) and (3.3) formally constitute a triangular array of the type $\{y_{tT} : t =$

$1, ..., T; T = 1, 2, ...\}$. Since this formulation is not essential in this paper, we will still use the notation $\{y_t\}$ to refer to this process. It has to be noted that, since $\alpha = E(y_t)(1-\rho)\phi(1)$, the process has no constant term if $\rho = 1$.

Given a sample from a process generated by (2.1) and (3.3), the analyst has to decide whether to estimate ρ or to impose the value $\rho = 1$. By the properties of least squares estimators it can be proved that the least squares estimator of ρ satisfies $\hat{\rho} = \rho + O_p\left\{T^{-(\beta+1)/2}\right\}$, whereas imposing unity has the property $1 = \rho + O(T^{-\beta})$. Then, for $\beta > 1$, the convergence rate when imposing unity is faster than estimating by least squares. This result helps to understand why processes with $\beta > 1$ are, for some purposes, better modeled in differences.

4 Properties of estimators in the overdifferenced process

4.1 Root-T consistency

Let $\{w_{t|p}\}$ be the true AR(p) process $\phi(B)w_{t|p} = a_t$. This process follows the Markovian representation $W_{t|p} = A_p W_{t-1|p} + U_{t,p}$. The $p \times p$ matrix A_p has the same structure than A_o with the coefficients $(\phi_1, \ldots, \phi_p)$ in the first row and $W_{t|p} = (w_{t|p}, ..., w_{t-p+1|p})'$. Then, from (2.3),

$$w_t = \phi^{-1}(B)\left\{1 - \frac{(1-\rho)B}{1-\rho B}\right\} a_t = w_{t|p} - \sum_{j=0}^{\infty} \psi_j (1-\rho) z_{t-1-j}, \tag{4.1}$$

where ψ_j are the coefficients of $\phi^{-1}(B)$, and $(1-\rho B)z_t = a_t$. Let us denote $\Gamma_{w|p} = E(W_{t|p}W'_{t|p})$ and $\gamma_{w|p} = E(W_{t|p}w_{t+1|p})$. We define the sampling autocovariances as $\hat{\Gamma}_{w|p} = (T-p-1)^{-1}\sum_{j=p+1}^{T-1} W_{j|p}W'_{j|p}$, $\hat{\gamma}_{w|p} = (T-p-1)^{-1}\sum_{j=p+1}^{T-1} W_{j|p}w_{j+1|p}$, and also make the following assumption:

A4.$E(\|\hat{\Gamma}^{-1}_{w|p}\|^{2k})$ $(k = 1, 2, ..., k_0)$ is bounded for all finite and sufficiently large T and some k_0.

The distance between the sampling second-order moments of w_t and $w_{t|p}$ is determined in the following theorem.

Theorem 1 *Let $\{w_t\}$ be the process (2.3) and let $w_1, ..., w_T$ be a sample from this process. Let ρ be defined as in 3.3 with $\beta \leq 1$, then*

(a) $\hat{\Gamma}_w = \hat{\Gamma}_{w|p} + O_p(T^{-\frac{1}{2}})$;

(b) $\hat{\gamma}_w = \hat{\gamma}_{w|p} + O_p(T^{-\frac{1}{2}})$.

See proof in Appendix B. Since $w_{t|p}$ is a stationary process, then $\hat{\gamma}_{w|p} = \gamma_{w|p} + O_p(T^{-\frac{1}{2}})$. Applying this result and theorem 1, the following corollary holds.

Corollary 1 *Assume the conditions of theorem 1 hold, then*

(a) $\hat{\Gamma}_w = \Gamma_{w|p} + O_p(T^{-\frac{1}{2}})$;

(b) $\hat{\gamma}_w = \gamma_{w|p} + O_p(T^{-\frac{1}{2}})$.

We can now prove root-T consistency of $\hat{\boldsymbol{\phi}}$. See proof in Appendix B.

Theorem 2 *Assume the conditions of theorem 1, then*

$$\hat{\boldsymbol{\phi}} = \boldsymbol{\phi} + O_p(T^{-\frac{1}{2}}).$$

4.2 Bias and mean squared error

Let $\hat{\boldsymbol{\phi}}_{|p}$ be the least squares estimator of $\boldsymbol{\phi}$ from a sample from a true AR(p) process. The bias and mean squared error (MSE) of this estimator, of a properly specified autoregression, have widely been investigated (see, for instance, Bhansali, 1981; Kunitomo & Yamamoto, 1985; Shaman & Stine, 1988; and references therein). Since the similarity between the estimator $\hat{\boldsymbol{\phi}}$, of the ARIMA$(p+1,1,1)$ misspecified as an AR(p), and $\hat{\boldsymbol{\phi}}_{|p}$ depends on the near nonstationarity hypothesis, we will express their differences in terms of ρ. The following theorems formulate the first and second order moments of the least squares estimator $\hat{\boldsymbol{\phi}}$ around the true parameter $\boldsymbol{\phi}$ as the respective moments of $\hat{\boldsymbol{\phi}}_{|p}$ plus an error term depending on ρ.

Theorem 3 *Assume A1 (with $s_o = 8$), A2, A3, and A4. Then*

$$E(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) = E(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi}) + O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}}\right\}. \tag{4.2}$$

The proof is in Appendix B. Since $(1-\rho)/(1+\rho) = O(T^{-\beta})$ and given that $E(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi}) = O(T^{-1})$ (see, for instance, Bhansali, 1981) we need a value $\beta > 2$ for the biases to be equal up to terms of order $O(T^{-1})$, whereas for root-T consistency we only need $\beta \geq 1$.

Theorem 4 *Assume A1 (with $s_o = 8$), A2, A3, and A4. Then*

$$E\{(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})'\} = E\{(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})'\} + O\left[max\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}, \frac{1-\rho}{1+\rho}\right\}\right].$$

See proof in Appendix B. We can see from this theorem that the MSE's are closer to each other than the biases. If ρ is such that $\beta > 1$ then both expressions for the MSE are equal up to terms $O(T^{-1})$.

5 Mean squared error of H-steps ahead prediction

In this section, we obtain the mean squared error of predicting y_{T+H} form $t = T$. The PMSE of a properly specified autoregression is (see, for instance, Kunitomo & Yamamoto, 1985)

$$\begin{aligned} \mathrm{PMSE}(\check{y}_{T+H}) &= \sigma^2 \sum_{h=0}^{H-1} (e'_{p+2} A_\alpha^h e_{p+2})^2 + \frac{\sigma^2}{T} \sum_{h=0}^{H-1} \sum_{k=0}^{H-1} (e'_{p+2} A_\alpha^h e_{p+2})(e'_{p+2} A_\alpha^k e_{p+2}) \\ &\quad \times \mathrm{tr}\left(A_\alpha^{H-1-h} \Gamma_y A'^{\,H-1-k}_\alpha \Gamma_y^{-1} \right) + O(T^{-3/2}). \end{aligned} \tag{5.1}$$

In order to compare the PMSE of the AR($p+1$) model (PMSE($\check{y}_{T+H}$)) with the PMSE of the misspecified ARIMA($p,1,0$) model (PMSE($\hat{y}_{T+H}$)) this expression is, however, inconvenient. We will rewrite the estimated H-steps ahead predictions in terms of their estimated increments ($\check{w}_t$ and $\hat{w}_t$, respectively). Hence, $\mathrm{PMSE}(\check{y}_{T+H}) = \sum_{h=1}^{H} \mathrm{PMSE}(\check{w}_{T+h}) + 2\sum_{h=1}^{H}\sum_{k=h+1}^{H} \mathrm{E}\{(w_{T+h} - \check{w}_{T+h})(w_{T+k} - \check{w}_{T+k})\}$, where $\check{w}_t = \check{y}_t - \check{y}_{t-1}$. A similar expression applies for PMSE($\hat{y}_{T+H}$) .

5.1 PMSE of the properly specified AR(p+1) predictor

Let $\check{A}_\alpha$ be the least squares estimator of A_α using the properly specified model (2.2). The estimated increment $\check{w}_{T+h}$ defined as a function of the estimated coefficients $\check{A}_\alpha$ is

$$\check{w}_{T+h} = e'_{p+2} \check{A}_\alpha^{h-1} (\check{A}_\alpha - I_{p+2}) Y_T, \tag{5.2}$$

where I_{p+2} is the identity matrix. The observed value w_{T+h} is

$$w_{T+h} = e'_{p+2} A_\alpha^{h-1} (A_\alpha - I_{p+2}) Y_T + L_h,$$

where $L_h = L_1 - L_2$, with $L_1 = \sum_{k=0}^{h-1} e'_{p+2} A_\alpha^k U_{T+h-k,p+2}$, and $L_2 = \sum_{k=1}^{h-1} e'_{p+2} A_\alpha^{k-1} U_{T+h-k,p+2}$.

The PMSE($\check{w}_{T+h}$) and $\mathrm{E}\{(\check{w}_{T+h} - w_{T+h})(\check{w}_{T+k} - w_{T+k})\}$ are shown in the following theorem (see proof in Appendix C). The assumptions about s_0 in theorems 5 and 6 are needed in order to apply the results of Kunitomo & Yamamoto (1985) in the proof of the theorems.

Theorem 5 *Let w_t follow (2.3), where $\rho = exp(-c/T^\beta)$ and $\beta > 1$. Assume A2, A3, A4, and A1 with $s_0 = 32$ when $h = 1,2$ and $s_0 = 16h$ when $h \geq 3$. Then*

$$\begin{aligned} PMSE(\check{w}_{T+h}) &= \sigma^2 \sum_{j=0}^{h-1} (e'_{p+2} A_1^j c_{p+2})^2 + \frac{\sigma^2}{T} \sum_{j=0}^{h-1}\sum_{k=0}^{h-1} (e'_p A_p^j e_p)(e'_p A_p^k e_p) \\ &\quad \times tr\left(A_\alpha^{h-1-j} \Gamma_y A'^{\,h-1-k}_\alpha \Gamma_y^{-1} \right) + O(T^{-\frac{3}{2}}), \end{aligned} \tag{5.3}$$

and, for $k \geq h$,

$$\begin{aligned} E\{(\check{w}_{T+h} - w_{T+h})(\check{w}_{T+k} - w_{T+k})\} &= \sigma^2 \sum_{i=0}^{h-1} (e'_{p+2} A_1^i c_{p+2})(e'_{p+2} A_1^{i+(k-h)} c_{p+2}) \\ + \frac{\sigma^2}{T} \sum_{n=0}^{k-1}\sum_{i=0}^{h-1} (e'_p A_p^n e_p)(e'_p A_p^i e_p) &\times tr\left(A_\alpha^{h-1-i} \Gamma_y A'^{\,k-1-n}_\alpha \Gamma_y^{-1} \right) + O(T^{-\frac{3}{2}}), \end{aligned} \tag{5.4}$$

where $c_{p+2} = (1, 0, ..., 0, 1)'$.

The terms on the right hand side of (5.3) and (5.4) have two components. The first component includes the variance of the prediction errors and the covariance between prediction errors at different horizons, respectively, of the noninvertible ARMA($p+1, 1$) process. The second component is the sampling error, due to the estimation of the $p+2$ parameters of the vector $\boldsymbol{\varphi}$.

5.2 PMSE of the overdifferenced ARIMA(p,1,0) predictor.

Assume that we predict w_{T+h} with the predictor derived from the estimated AR(p), that is $\hat{w}_{T+h} = e_p' \hat{A}_p^h W_T$, where $\hat{A}_p$ is the least squares estimator of A_p. Then

$$\hat{w}_{T+h} = e_p' A_p^h W_T + e_p'(\hat{A}_p^h - A_p^h) W_T = E(w_{T+h|p}|T) + e_p'(\hat{A}_p^h - A_p^h) W_T.$$

The true value w_{T+h} is, from (2.4), $w_{T+h} = e_{p+2}' A_1^h Z_T + L_h = E(w_{T+h}|T) + L_h$. Then the h-steps ahead prediction error is $(w_{T+h} - \hat{w}_{T+h}) = L_h - e_p'(\hat{A}_p^h - A_p^h) W_T - v_t$, where, by (4.1),

$$v_t = E(w_{T+h} - \hat{w}_{T+h|p}|T) = \sum_{j=h-1}^{\infty} \psi_j (1-\rho) z_{T+h-1-j} + \sum_{j=0}^{h-2} \psi_j (1-\rho) \rho^{h-1-j} z_T. \tag{5.5}$$

The following theorem gives an approximation of order $o(T^{-1})$ of the expectation of the lead-h mean squared prediction error (see proof in Appendix C).

Theorem 6 *Let w_t follow (2.3), where $\rho = exp(-c/T^{\beta})$ and $\beta > 1$. Assume A2, A3, A4, and A1 with $s_0 = 32$ when $h = 1, 2$ and $s_0 = 16h$ when $h \geq 3$. Then*

$$\begin{aligned} PMSE(\hat{w}_{T+h}) &= \sigma^2 \sum_{k=0}^{h-1} (e_{p+2}' A_1^k c_{p+2})^2 + \frac{\sigma^2}{T} \sum_{j=0}^{h-1} \sum_{k=0}^{h-1} (e_p' A_p^j e_p)(e_p' A_p^k e_p) \\ & \quad \times tr\left(A_p^{h-1-j} \Gamma_{w|p} A_p'^{\,h-1-k} \Gamma_{w|p}^{-1} \right) + o(T^{-1}), \end{aligned} \tag{5.6}$$

and, for $k \geq h$,

$$\begin{aligned} E\{(\hat{w}_{T+h} - w_{T+h})(\hat{w}_{T+k} - w_{T+k})\} &= \sigma^2 \sum_{i=0}^{h-1} (e_{p+2}' A_1^i c_{p+2})(e_{p+2}' A_1^{i+(k-h)} c_{p+2}) \\ + \frac{\sigma^2}{T} \sum_{n=0}^{k-1} \sum_{i=0}^{h-1} (e_p' A_p^n e_p)(e_p' A_p^i e_p) \times tr\left(A_p^{h-1-i} \Gamma_{w|p} A_p^{k-1-n} \Gamma_{w|p} p^{-1} \right) + o(T^{-1}), \end{aligned} \tag{5.7}$$

where $c_{p+2} = (1, 0, ..., 0, 1)'$.

The terms on the right hand side of (5.6) and (5.7) have two components. The first one, the variance of prediction errors and their covariance between different horizons of the true ARIMA($p+1, 1, 1$) process, is the same than in theorem 5. The second one is the sampling error due to the estimation of the p parameters $\boldsymbol{\phi}$, in contrast with the estimation of the $p+2$ parameters of the AR($p+1$) model. It should be observed that this second component differ from the one on the previous subsection only in the elements inside the trace operators.

6 Comparing prediction accuracy

In this section, we compare the PMSE's found in the last section for the two models. We prove that, under the assumption of near nonstationarity exposed in (3.3), with $\beta > 1$, overdifferencing may produce lower PMSE (to terms of small order). The expressions in theorem 5 and theorem 6 reveal that the only difference between PMSE($\check{y}_{T+H}$) and PMSE($\hat{y}_{T+H}$) is in the elements inside the trace operators. These traces can be compared using the two following lemmas: lemma 1 compares such a trace in processes with and without constant term; lemma 2 compares the trace in nearly nonstationary processes with no constant term and the overdifferenced one. The proofs of these lemmas can be found in Appendix D.

Lemma 1 *Let y_t follow process (2.1). Then*

$$tr\left(A_\alpha^i \Gamma_y A_\alpha'^j \Gamma_y^{-1}\right) = 1 + tr\left(A_o^i \Gamma_{\tilde{y}} A_o'^j \Gamma_{\tilde{y}}^{-1}\right).$$

Lemma 2 *Let y_t follow process (2.1) with $\rho = exp(-c/T^\beta)$ and $\beta > 1$. Then*

$$tr(A_o^i \Gamma_{\tilde{y}} A_o'^j \Gamma_{\tilde{y}}^{-1}) = \rho^{i+j} + tr\left(A_p^i \Gamma_{w|p} A_p^j \Gamma_{w|p}^{-1}\right) + o(T^{-1}).$$

Now we can prove the advantage of overdifferencing when the process is nearly nonstationary.

Theorem 7 *Let y_t follow process (2.1) with $\rho = exp(-c/T^\beta)$ and $\beta > 1$, and let the conditions of theorems 5 and 6 hold. Then, for $H \geq 1$,*

$$PMSE(\check{y}_{T+H}) - PMSE(\hat{y}_{T+H}) = \nu_H + o(H^2 T^{-1}),$$

where

$$\nu_H = \frac{\sigma^2}{T}\left(\sum_{h=1}^{H}\sum_{j=0}^{h-1}\psi_j\right)^2 + \frac{\sigma^2}{T}\left(\sum_{h=1}^{H}\sum_{j=0}^{h-1}\psi_j \rho^{h-1-j}\right)^2 > 0, \tag{6.1}$$

with $\psi_j = (e_p' A_p^j e_p)$, $(j = 1, ..., H)$.

The proof is a direct application of lemma 1 and lemma 2 to the differences between (5.3) and (5.6) and between expression (5.7) and (5.4).

Expression (6.1) shows that the advantage of the overdifferenced model can be decomposed into two parts. The first term at the right side of (6.1) is the result of applying lemma 1 and, therefore, is due to the MSE of estimating the constant term α in the AR$(p+1)$ model. The second term is the result of applying lemma 2 and, then, is due to the MSE of estimating an extra parameter in the AR$(p+1)$. Thus, the superior forecasting performance of the model ARIMA$(p,1,0)$ is due to its more parsimonious representation. For $H=1$ the difference is $2\sigma^2/T$ if a constant is needed, and σ^2/T if $\alpha=0$ and no constant is estimated. This result is similar to that of Ledolter & Abraham (1981) for overspecified models, where they state that each unnecessary estimated parameter increases the one-step ahead PMSE by σ^2/T.

Although these results are applicable to a general stationary autoregression, it is interesting to analyze the AR(1) case. First, its simplicity avoids the use of some asymptotic approximations. Second, the results will not be affected by any other root, as shown in (3.2), and they can be considered as a neutral benchmark. The PMSE of the proper predictor in this case can be evaluated with (5.1), whereas the PMSE in the overdifferenced model is easily evaluated using as predictor a random walk. The following remarks summarize the results for both the AR(1) case with no intercept (AR(1)) and with intercept (AR(1,μ)).

Remark 1. Let y_t follow the process $y_t = \rho y_{t-1} + a_t \;\;, |\rho| < 1$. Then PMSE $(\check{y}_{T+H})-$PMSE$(\hat{y}_{T+H}) = \nu_{H|\text{ AR}(1)} + o(H^2T^{-\frac{3}{2}})$, where

$$\nu_{H|\text{AR}(1)} = \sigma^2 \left\{ \frac{H^2\rho^{2(H-1)}}{T} - \frac{(1-\rho^H)^2}{1-\rho^2} \right\}. \tag{6.2}$$

Table 1 shows the values of ρ that make $\nu_{H|\text{AR}(1)} = 0$. Larger values will produce $\nu_{H|\text{AR}(1)} > 0$. These values of ρ increase with H. Therefore, as the horizon grows, the process needs to be closer to the unit root in order to get some gain when differencing. The advantage of overdifferencing tends, then, to decrease when the horizon is large. It can also be seen that as $H \to \infty$ the limit of (6.2) is negative. Then, the advantage of the overdifferenced predictor eventually disappears. If ρ is close enough to unity, this will happen at a horizon of no practical interest. This result has an interpretation in terms of the time reversibility of the true process. Since the process is stationary, its long term prediction is the unconditional mean, which in this case is known. Therefore, the AR(1) predictor will forecast the long term with no error, whereas the random walk will not. Manipulating (6.2), we can conclude that, up to terms of small order, overdifferencing can produce better forecasts if

$$\rho > \exp\left(-\frac{2}{T+4H}\right). \tag{6.3}$$

This expression can be approximated, omitting the influence of H, as $\rho > \exp(-2/T)$. This value of $c=2$ agrees with the empirical work of Stock (1996).

Table 1: Values of ρ to obtain $\nu_{H|\mathrm{AR}(1)} = 0$ and $\nu_{H|\mathrm{AR}(1,\mu)} = 0$.

	AR(1)					AR(1,μ)				
	Horizon					Horizon				
T	1	2	5	10	20	1	2	5	10	20
25	0.923	0.937	0.940	0.951	0.963	0.852	0.862	0.881	0.898	0.913
50	0.961	0.965	0.966	0.970	0.976	0.923	0.926	0.932	0.940	0.948
75	0.974	0.976	0.976	0.978	0.982	0.948	0.949	0.953	0.957	0.962
100	0.980	0.981	0.982	0.983	0.985	0.961	0.962	0.964	0.966	0.970
150	0.987	0.987	0.987	0.988	0.989	0.974	0.974	0.975	0.976	0.978
300	0.993	0.994	0.994	0.994	0.994	0.987	0.987	0.987	0.988	0.988

Remark 2. Let y_t follow the process $y_t = \alpha + \rho y_{t-1} + a_t \quad , |\rho| < 1$. Then PMSE $(\check{y}_{T+H})$ − PMSE$(\hat{y}_{T+H}) = \nu_{H|\mathrm{AR}\ (1,\mu)} + o(H^2 T^{-\frac{3}{2}})$, where

$$\nu_{H|\mathrm{AR}\ (1,\mu)} = \sigma^2 \left\{ \frac{H^2 \rho^{2(H-1)}}{T} + \frac{(1-\rho^H)^2}{T(1-\rho)^2} - \frac{(1-\rho^H)^2}{1-\rho^2} \right\}. \tag{6.4}$$

Table 1 shows the values of ρ that make $\nu_{H|\mathrm{AR}(1,\mu)} = 0$. From (6.4) it can be verified that the overdifferenced predictor produces better forecasts, up to terms of small order, if

$$\rho > \exp\left(-\frac{4}{T+4H}\right), \tag{6.5}$$

that can be simplified as $\rho > \exp(-4/T)$. In this case, the limit of (6.4) as $H \to \infty$ is still positive if $\rho > \exp(-2/T)$.

7 A simulation study

In this section, we illustrate the preceding results with a simulation exercise. We consider three different AR(2) models: M1: $(1-0.5B)(1-\rho B)y_t = 10 + a_t$; M2: $(1-0.5B)(1-\rho B)y_t = a_t$; and M3: $(1+0.8B)(1-\rho B)y_t = 10 + a_t$, with $\rho = 0.9,\ 0.92,\ 0.94,\ 0.96,\ 0.98,\ 0.99$. Sample sizes are $T = 50,\ 100$. Real series usually have non-zero mean, and models M1 and M3 can illustrate the consequences of overdifferencing in such series. Also, model M2 can arise when in doubt about a second difference.

An important aspect in the simulation exercise is the possibility of obtaining an explosive estimated predictor. There are two main reasons to avoid these explosive situations. Firstly, they are of limited

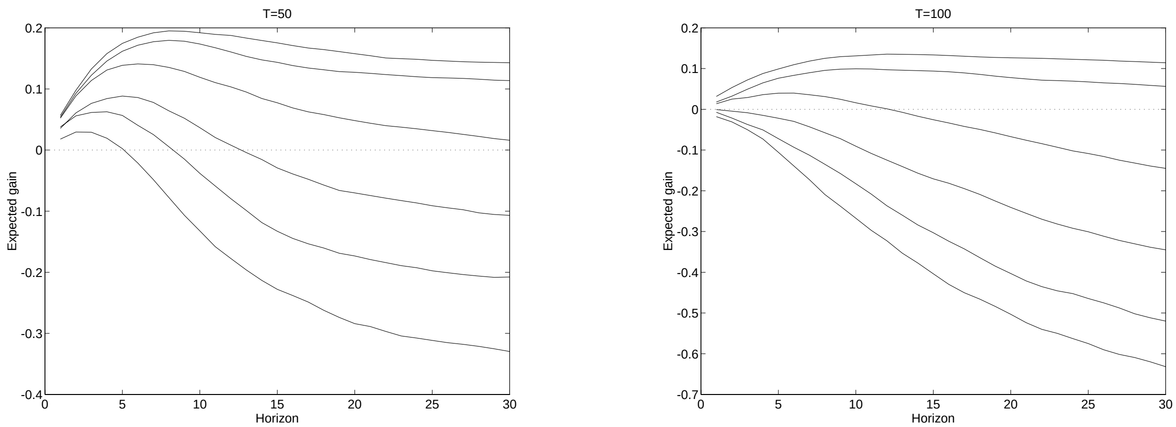

Figure 1: $\{V_y(H) - V_w(H)\}/V_y(H)$ of model M1 for horizon $H = 1, ..., 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

practical interest. A typical situation where a practitioner has doubts about differencing, for forecasting purposes, deals mainly with estimated ρ close to, but lower than unity. Second, the explosive nature of the predictions generated with a predictor with $\hat{\rho} > 1$ produces an excessive influence on the averages resulting from the simulations, because explosive estimated predictor is easily worse than its overdifferenced counterpart, especially at long term. Unreported simulations show that very few explosive estimated predictors can have an extremely high influence in the computations, giving a too optimistic representation of the effect of overdifferencing. Therefore, in order to obtain a clearer picture of what can be expected from overdifferencing in a real situation, we have considered only those replications whose estimated roots where outside the unit circle. The percentage of rejected replications is low. For instance, if $\rho = 0.98$ and $T = 100$ this is 1%, and with $T = 50$ it is 2.7%.

In each replication, we generate a random sample of the process of size $500 + T + 30$ with random noise $a_t \sim N(0, 1)$. The first 500 observations were ignored to avoid the effect of initial values, and the last 30 were used to evaluate the prediction error. By averaging the predicting squared errors of 20000 valid replications we obtain $V_y(H)$ and $V_w(H)$ as the sampling estimation of the PMSE of forecasting y_{T+H} using the forecasts generated by the correct AR(2) model or the overdifferenced ARIMA(1,1,0) model respectively. Figures 1 to 3 show the ratio $\{V_y(H) - V_w(H)\}/V_y(H)$ for M1 to M3 as a function of T and ρ. This ratio represents the empirical expected gain (or loss if negative) of overdifferencing at each horizon.

These figures reveal that, as expected from the theoretical results, there are situations where overdifferencing outperformed the true model. The expected gain increases with the size of ρ and decreases

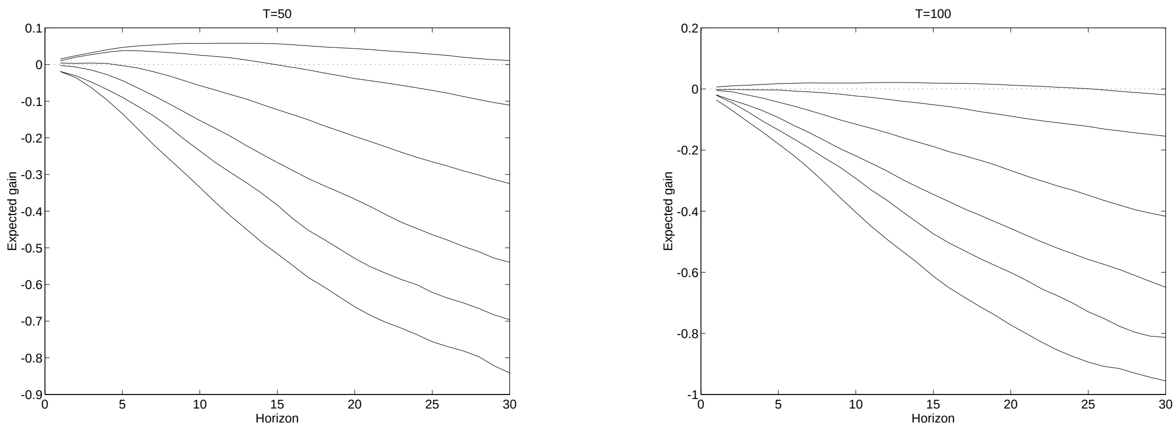

Figure 2: $\{V_y(H) - V_w(H)\}/V_y(H)$ of model M2 for horizon $H = 1, ..., 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

with T. Also, in agreement with equation (3.2), the gain is larger in the model with positive second root (M1) than in the model with negative root (M3). The gain substantially decreases if $\alpha = 0$ (M2).

The main feature of these figures is the divergence of the curves as the horizon increases. In the very short term, the difference between the two predictors is very small, even negligible. Nevertheless, in the medium or long term the gain or loss can be important. The risk of falling into an important loss if ρ is not large enough can, however, be diminished if some efficient rule to decide about differencing is used. A second important aspect of these figures is that in the long run ($H \gg T^{1/2}$) the gain decreases and can be negative. Also, as proved in the last section, the gain in the model with no constant always disappears at sufficiently large H.

Figures 4 and 5 show the absolute values of $V_y(H)$ and $V_w(H)$ for selected values of ρ. These figures also contain the population PMSE of the process (dotted lines). These population values can be obtained from the first term on the right side of expression (5.1). The distance from these population curves to each solid line is the PMSE due to the estimation of the unknown parameters. It can be seen that the sampling variability of the nondifferenced predictor (line with symbol +) increases notably when the number of parameters increases (model M1 and M3 with respect to M2). This increment of the PMSE due to the estimation of the parameters makes that the overdifferenced predictor (line with symbol o) can outperform its competitor when the process approach nonstationarity.

It can be seen that the theoretical results accurately explain this finite sample performance. Since results depend mainly on the size of the roots rather than on its number, it is reasonable to foresee

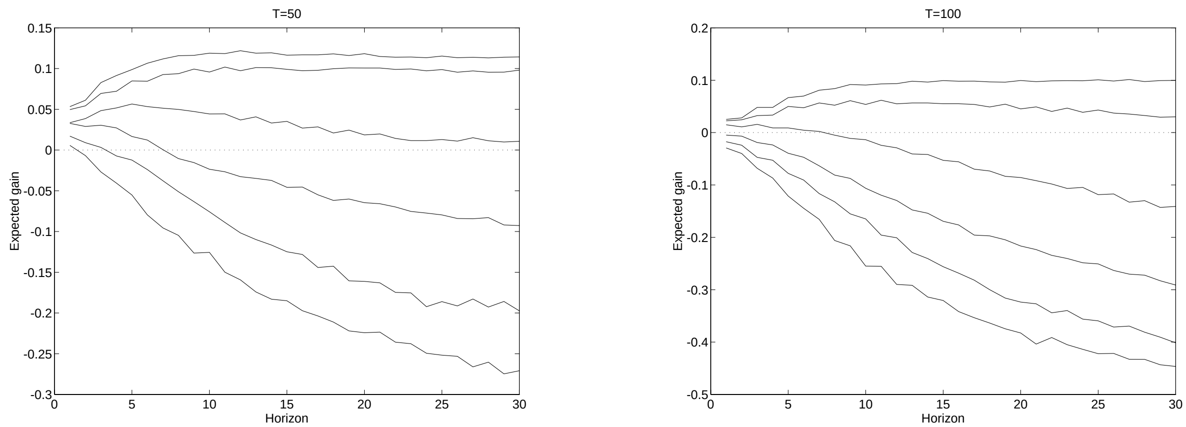

Figure 3: $\{V_y(H) - V_w(H)\}/V_y(H)$ of model M3 for horizon $H = 1, \ldots, 30$ and sample size T. The values of ρ are (from down to top): 0.90, 0.92, 0.94, 0.96, 0.98 , 0.99.

similar conclusions in larger autoregressions.

Acknowledgment

This work has been partially supported by DGES, grant PB96-0111, and Cátedra BBV de Métodos para la Mejora de la Calidad. The authors would like to thank Ngai Hang Chan, George Tiao, Mike Wiper, and the participants of the NBER/NSF Time Series Seminar, Duke University, 1997, for helpful discussions and suggestions on this work. We also thank the referees and the editor for their valuable and constructive comments.

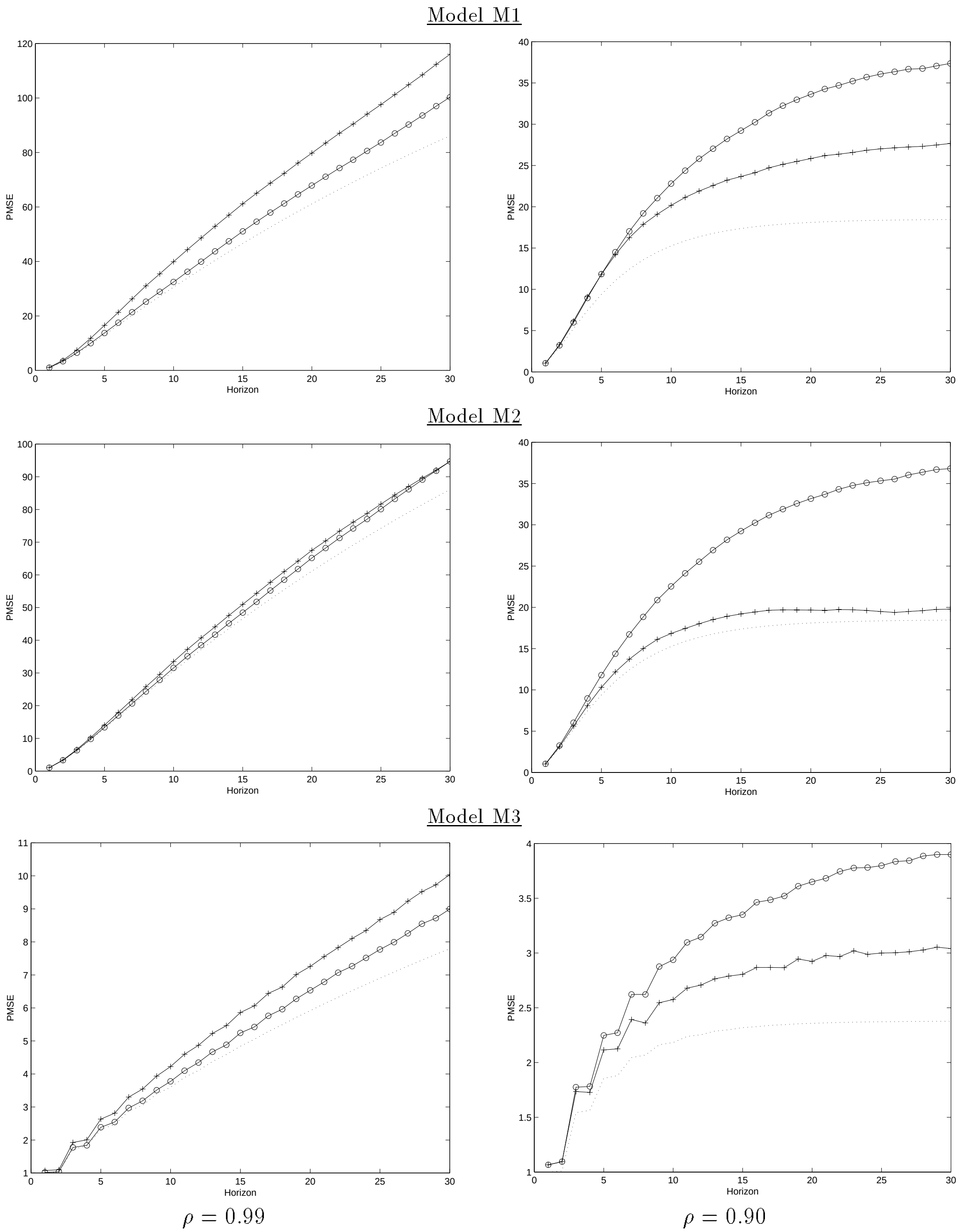

Figure 4: Values of V_y (line with symbol $+$), V_w (line with symbol o), and population PMSE (dotted line). Sample size $T = 50$.

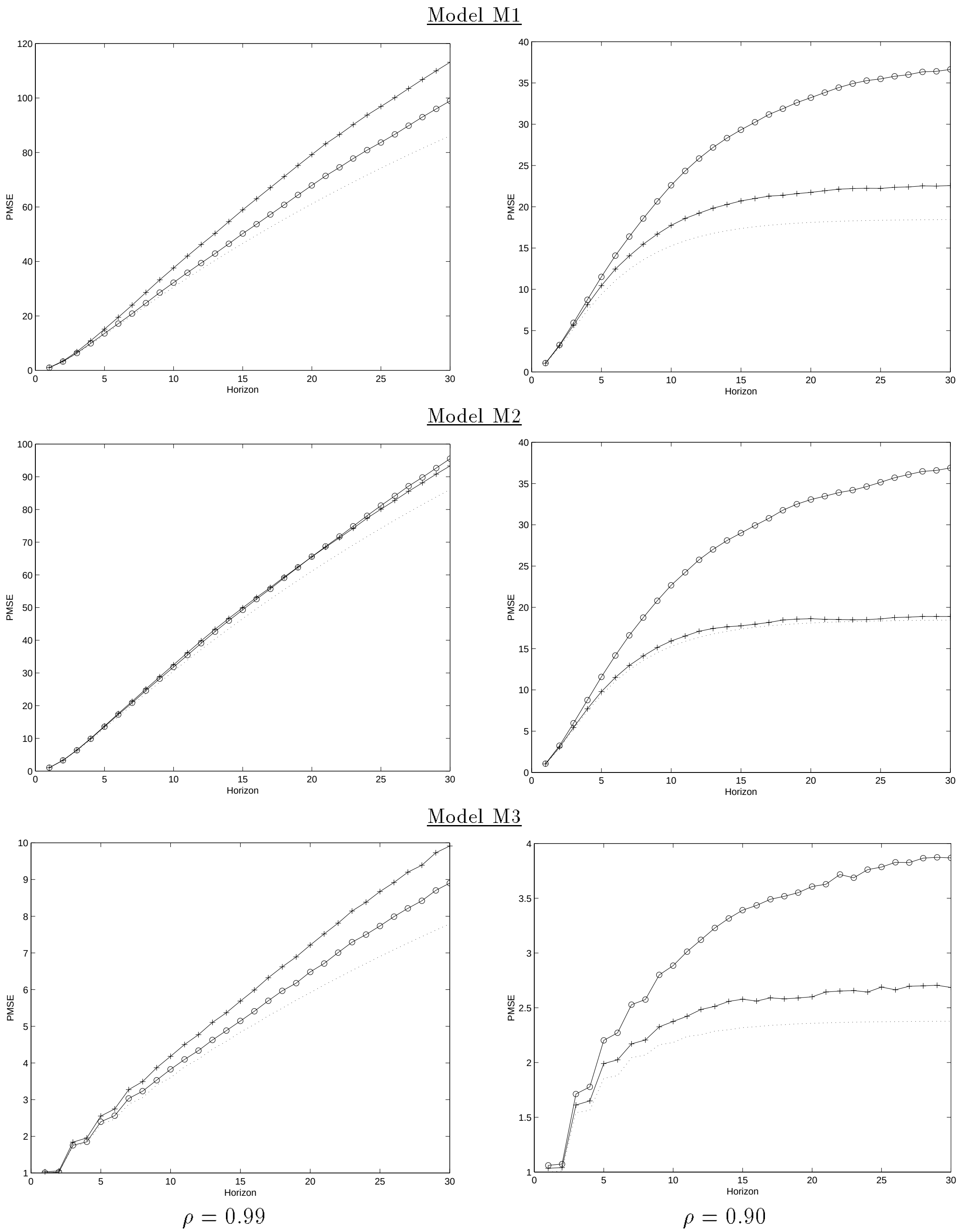

Figure 5: Values of V_y (line with symbol +), V_w (line with symbol o), and population PMSE (dotted line). Sample size $T = 100$.

APPENDIX

A Lemmas

We present some lemmas used for the proof of theorems in subsequent sections. For an arbitrary $p \times 1$ vector x and a $p \times p$ matrix M, let $\|x\| = (x'x)^{1/2}$ be the Euclidean norm of x and $\|M\| = \sup_{\|x\| \leq 1}(x'M'Mx)^{1/2}$ be the matrix norm of M.

Lemma A.1 *Assume A1 and A2, with $s_o = 2k$ and $k \geq 1$. Then, as $T \to \infty$,*

$$E(\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^k) = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}\right\},$$

and

$$E(\|\hat{\gamma}_w - \hat{\gamma}_{w|p}\|^k) = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}\right\}. \tag{A.1}$$

Proof: Let m_{ij} be a generic element of M. Since $E(\|M\|^k) = O\left\{\max_{i,j} E(|m_{ij}|^k)\right\}; i,j = 1,..,p;$ and by Minkowski's inequality, $E(\|\hat{\Gamma}_w - \hat{\Gamma}_{w|p}\|^k) = O\left(\max_{t,s} E|w_t w_{t-s} - w_{t|p}w_{t-s|p}|^k\right)$. A similar result applies to (A.1). Using the decomposition (4.1), and by Minkowski's inequality,

$$E|w_t w_{t-s} - w_{t|p}w_{t-s|p}|^k \leq \left\{\left(E|w_{t|p}r_{t-s}|^k\right)^{\frac{1}{k}} + \left(E|w_{t-s|p}r_t|^k\right)^{\frac{1}{k}} + \left(E|r_t r_{t-s}|^k\right)^{\frac{1}{k}}\right\}^k, \tag{A.2}$$

where $r_t = \sum_{j=0}^{\infty} \psi_j(1-\rho)z_{t-1-j}$. By Hölders' inequality, $E|w_{t|p}r_{t-s}|^k \leq \left(E|w_{t|p}|^{2k}E|r_{t-s}|^{2k}\right)^{\frac{1}{2}}$. Also, by assumption A1, $E|w_{t|p}|^{2k} = O(1)$. Similarly, $E|r_{t-s}|^{2k} \leq \left\{\sum_{j=0}^{\infty}|\psi_j(1-\rho)|\left(E|z_{t-s-1-j}|^{2k}\right)^{\frac{1}{2k}}\right\}^{2k}$, where it can be verified, that

$$E|z_t|^{2k} \leq E\left(\sum_{j=0}^{\infty}\left|\rho^{2j}a_{t-j}^2\right|\right)^k \leq \left\{\sum_{j=0}^{\infty}\left(E\left|\rho^{2j}a_{t-j}^2\right|^k\right)^{\frac{1}{k}}\right\}^k = O\left\{\left(\frac{1}{1-\rho^2}\right)^k\right\}.$$

Therefore,

$$E|r_{t-s}|^{2k} = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^k\right\}, \tag{A.3}$$

and hence,

$$E|w_{t|p}r_{t-s}|^k = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}\right\}.$$

A similar result applies to the second term in (A.2). The third term in (A.2) can also be solved following the previous arguments. It can be shown that

$$E|r_t r_{t-s}|^k = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^k\right\}.$$

Applying these results to (A.2) proves the lemma. □

Lemma A.2 *Assume A1, A2, A3, and A4, with $s_0 = 2k$. Then, as $T \to \infty$,*

$$E(\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^k) = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}\right\}.$$

Proof: It can be verified that $\|\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}\|^k = \|\hat{\Gamma}_w^{-1}(\hat{\Gamma}_w - \hat{\Gamma}_{w|p})\hat{\Gamma}_{w|p}^{-1}\|^k$. By Hölders' inequality and lemma A.1 the result follows. □

Lemma A.3 *Assume A1, A2, A3, and A4, with $s_0 = 4k$. Then, as $T \to \infty$,*

$$E(\|\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p}\|^k) = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}\right\}, \tag{A.4}$$

$$E(\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|^k) = O\left[\max\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{k}{2}}, T^{-\frac{k}{2}}\right\}\right]. \tag{A.5}$$

Proof: The estimator $\hat{\boldsymbol{\phi}}$ can be expressed as

$$\hat{\boldsymbol{\phi}} = (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p} + \hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p}) + \hat{\boldsymbol{\phi}}_{|p},$$

where $\hat{\boldsymbol{\phi}}_{|p} = \hat{\Gamma}_{w|p}^{-1}\hat{\gamma}_{w|p}$. By Minkowski's inequality we obtain

$$\begin{aligned} E(\|\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p}\|^k) \;\leq\; & \Big(\left[E\{\|(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|^k\}\right]^{\frac{1}{k}} \\ & + \left[E\{\|(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p}\|^k\}\right]^{\frac{1}{k}} + \left[E\{\|\hat{\Gamma}_{w|p}^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p})\|^k\}\right]^{\frac{1}{k}}\Big)^k. \end{aligned}$$

By Hölder's inequality and applying lemmas A.1 and A.2 expression (A.4) holds. In order to prove (A.5) we use the decomposition $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} = \Gamma_{w|p}^{-1}(\hat{\gamma}_w - \gamma_{w|p}) + (\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1})\hat{\gamma}_w$, and also the decompositions $\hat{\gamma}_w - \gamma_{w|p} = (\hat{\gamma}_w - \hat{\gamma}_{w|p}) + (\hat{\gamma}_{w|p} - \gamma_{w|p})$ and $\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1} = (\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1}) + (\hat{\Gamma}_{w|p}^{-1} - \Gamma_{w|p}^{-1})$. Applying that $E(\|\hat{\gamma}_{w|p} - \gamma_{w|p}\|^{2k}) = O(T^{-k})$ and $E(\|\hat{\Gamma}_{w|p}^{-1} - \Gamma_{w|p}^{-1}\|^{2k}) = O(T^{-k})$ (see, for instance, lemma 3.3 of Bhansali, 1981), and using the same arguments as before, completes the result. □

B Proofs of results in section 3

Proof of theorem 1:

Since $E(z_t^2) = \sigma^2/(1-\rho^2)$, and by Chebyshev's Inequality, we obtain $z_t = O_p\left\{(1-\rho^2)^{-\frac{1}{2}}\right\}$. Hence,

$$r_t = O_p\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}}\right\}. \tag{B.1}$$

Since $(1-\rho)/(1+\rho) = O(T^{-\beta})$, then $w_t = w_{t|p} + o_p\left(T^{-\frac{1}{2}}\right)$.

The elements of $\hat{\Gamma}_w$ and $\hat{\gamma}_w$ can be decomposed as

$$\begin{aligned}\frac{\sum_{j=p+1}^{T-1} w_{j-t}w_{j-s}}{T-p-1} &= \frac{\sum_{j=p+1}^{T-1} w_{j-t|p}w_{j-s|p}}{T-p-1} - \frac{\sum_{j=p+1}^{T-1} w_{j-t|p}r_{j-s}}{T-p-1} - \frac{\sum_{j=p+1}^{T-1} w_{j-s|p}r_{j-t}}{T-p-1} \\ &\quad + \frac{\sum_{j=p+1}^{T-1} r_{j-t}r_{j-t}}{T-p-1}.\end{aligned}$$

Applying (B.1) and the result that $w_{t|p} = O_p(1)$, it can be verified that

$$\frac{\sum_{j=p+1}^{T-1} w_{j-t}w_{j-s}}{T-p-1} = \frac{\sum_{j=p+1}^{T-1} w_{j-t|p}w_{j-s|p}}{T-p-1} + o_p(T^{-\frac{1}{2}})$$

and the theorem follows. □

Proof of theorem 2: Using the decomposition $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} = \Gamma_{w|p}^{-1}(\hat{\gamma}_w - \gamma_{w|p}) + (\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1})\hat{\gamma}_w$, and by stationarity of $\{w_{t|p}\}$, we have $\Gamma_{w|p}^{-1} = O(1)$. Also, if $\hat{\Gamma}_w^{-1}$ exists, we have $(\hat{\Gamma}_w^{-1} - \Gamma_{w|p}^{-1}) = \hat{\Gamma}_w^{-1}(\Gamma_{w|p} - \hat{\Gamma}_w)\Gamma_{w|p}^{-1}$. Therefore, applying corollary 1, $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} = O_p(T^{-1/2})$. □

Proof of theorem 3: It can be verified that $E(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p}) = E\{(\hat{\Gamma}_w^{-1} - \hat{\Gamma}_{w|p}^{-1})\hat{\gamma}_{w|p}\} + E\{\hat{\Gamma}_w^{-1}(\hat{\gamma}_w - \hat{\gamma}_{w|p})\}$. Applying Hölders' inequality and lemmas A.2 and A.1 the theorem follows. □

Proof of theorem 4: We can decompose

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\phi}}) &= \mathrm{MSE}(\hat{\boldsymbol{\phi}}_{|p}) + E\left\{(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})'\right\} \\ &\quad + E\left\{(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})'\right\} + E\left\{(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})'\right\}.\end{aligned}$$

Since $\|M\| \le \sqrt{\mathrm{tr}(M'M)}$, and applying lemma A.3,

$$E\left\{\|(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})'\|\right\} \le E(\|\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p}\|^2) = O\left(\frac{1-\rho}{1+\rho}\right).$$

Analogously, and applying the result that $E\left(\|\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi}\|^2\right) = O(T^{-1})$ (see, for instance, Bhansali, 1981), it can be verified that

$$E\left\{\|(\hat{\boldsymbol{\phi}}_{|p} - \boldsymbol{\phi})(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_{|p})'\|\right\} = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}\right\},$$

$$E\left\{\|(\hat{\boldsymbol{\phi}}-\hat{\boldsymbol{\phi}}_{|p})(\hat{\boldsymbol{\phi}}_{|p}-\boldsymbol{\phi})'\|\right\} = O\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}\right\},$$

and the theorem follows. □

C Proofs of results in section 4:

Proof of theorem 5: The Taylor expansions of $\check{A}_\alpha^h$ and $\check{A}_\alpha^{h-1}$ around A_α are

$$\check{A}_\alpha^k = A_\alpha^k + \sum_{j=0}^{k-1} A_\alpha^j(\check{A}_\alpha - A_\alpha)A_\alpha^{k-1-j} + O_p(T^{-1});\ k = h, h-1.$$

Then, using $\sum_{j=0}^{h-2} A_\alpha^j(\check{A}_\alpha - A_\alpha)A_\alpha^{h-2-j} = \sum_{j=1}^{h-1} A_\alpha^{j-1}(\check{A}_\alpha - A_\alpha)A_\alpha^{h-1-j}$, and given that $(\check{A}_\alpha - A_\alpha) = e_{p+2}(\check{\boldsymbol{\varphi}} - \boldsymbol{\varphi})'$, we have

$$\begin{aligned} E(\check{w}_{T+h} - w_{T+h})^2 &= E(L_1 - L_2)^2 + E(C_{h,1}'Y_TY_T'C_{h,1}) + E(C_{h,2}'Y_TY_T'C_{h,2}) \\ &\quad + E(C_{h,1}'Y_TY_T'C_{h,2}) + E(C_{h,2}'Y_TY_T'C_{h,1}) + O(T^{-\frac{3}{2}}), \end{aligned} \tag{C.1}$$

where $C_{h,1}' = e_{p+2}'A_\alpha^0 e_{p+2}(\check{\boldsymbol{\varphi}} - \boldsymbol{\varphi})'A_\alpha^{h-1}$, and $C_{h,2}' = \sum_{j=1}^{h-1} e_{p+2}'A_\alpha^{j-1}(A_\alpha - I_{p+2})e_{p+2}(\check{\boldsymbol{\varphi}} - \boldsymbol{\varphi})'A_\alpha^{h-1-j}$, and where we have used the result that $E(\|\check{A}_\alpha - A_\alpha\|^k) = O(T^{-\frac{k}{2}})$ (see, for instance, Bhansali, 1981, or Kunitomo & Yamamoto, 1985) .

If we denote the k-th coefficient of $\varphi(B)^{-1}$ by $\psi_{k[\mathrm{AR}(p+1)]}$ and the k-th coefficient of $\varphi(B)^{-1}(1-B)$ by $\psi_{k[\mathrm{ARMA}(p+1,1)]}$, then $e_{p+2}'A_\alpha^{k-1}(A_\alpha - I_{p+2})\,e_{p+2} = \psi_{h[\mathrm{AR}(p+1)]} - \psi_{k-1[\mathrm{AR}(p+1)]} = \psi_{k[\mathrm{ARMA}(p+1,1)]} = e_{p+2}'A_1^k c_{p+2}$, and hence,

$$E(L_h^2) = E\left\{(L_1 - L_2)^2\right\} = \sigma^2 \sum_{k=0}^{h-1}(e_{p+2}A_1^h c_{p+2})^2. \tag{C.2}$$

Since the effect of the dependence between Y_T and $\check{\boldsymbol{\varphi}}$ in the PMSE is $O(T^{-\frac{3}{2}})$ (Kunitomo & Yamamoto, 1985) and applying that $\mathrm{MSE}(\check{\boldsymbol{\varphi}}) = \sigma^2\Gamma_y^{-1}/T + O(T^{-\frac{3}{2}})$, we find

$$\begin{aligned} E(C_{h,2}'Y_TY_T'C_{h,2}) &= \frac{\sigma^2}{T}\sum_{j=1}^{h-1}\sum_{k=1}^{h-1}\left(e_{p+2}'A_1^{j-1}c_{p+2}\right)\left(e_{p+2}'A_1^{k-1}c_{p+2}\right) \\ &\quad \times \mathrm{tr}\left(A_\alpha^{h-1-j}\Gamma_y A_\alpha'^{\,h-1-k}\Gamma_y^{-1}\right) + O(T^{-\frac{3}{2}}). \end{aligned}$$

Applying the same arguments to the remaining terms of (C.1) we obtain

$$\begin{aligned} E[(\check{w}_{T+h} - w_{T+h})^2] &= \sigma^2\sum_{k=0}^{h-1}(e_{p+2}'A_1^k c_{p+2})^2 + \frac{\sigma^2}{T}\sum_{j=0}^{h-1}\sum_{k=0}^{h-1}(e_{p+2}'A_1^j c_{p+2})(e_{p+2}'A_1^k c_{p+2}) \\ &\quad \times \mathrm{tr}\left(A_\alpha^{h-1-j}\Gamma_y A_\alpha'^{\,h-1}\Gamma_y^{-1}\right) + O(T^{-\frac{3}{2}}). \end{aligned}$$

If we denote the k-th coefficient of $\phi(B)^{-1}$ by $\psi_{k[\mathrm{AR}(p)]}$, then $\psi_{k[\mathrm{ARMA}(p+1,1)]} = \psi_{k[\mathrm{AR}(p)]} + O(1-\rho)$ and, therefore, $e'_{p+2}A_1^k c_{p+2} = e'_p A_p^k e_p + O(1-\rho)$. Then, if $\beta > 1$, expression (5.3) holds. Similarly, using the previous arguments, the proof of (5.4) follows. $\square$

Proof of theorem 6: The expectation of the square of $w_{T+h} - \hat{w}_{T+h}$ is

$$\begin{aligned} E\{(w_{T+h} - \hat{w}_{T+h})^2\} &= E(L_h^2) + E\left\{e'_p(\hat{A}_p^h - A_p^h)W_T W'_T(\hat{A}_p^h - A_p^h)'e_p\right\} \\ &\quad + E(v_T^2) + 2E\left\{e'_p(\hat{A}_p^h - A_p^h)W_T v_T\right\}, \end{aligned} \tag{C.3}$$

where the term $E(L_h^2)$ is the same than (C.2). Applying (A.3) with $k=1$ and Hölders' inequality, then $E(v_T^2) = o(T^{-1})$. In order to solve the remaining terms of (C.3), we will use a Taylor expansion of $\hat{A}_p$ around A_p. The magnitude of the remainder term is determined by the root-T consistency of $\hat{A}_p$. Then

$$\begin{aligned} \hat{A}_p^h &= A_p^h + \sum_{j=0}^{h-1} A_p^j(\hat{A}_p - A_p)A_p^{h-1-j} \\ &\quad + \sum_{j=1}^{h-1}\left\{\sum_{k=0}^{j-1} A_p^k(\hat{A}_p - A_p)A_p^{j-1-k}\right\} \times (\hat{A}_p - A_p)A_p^{h-1-j} + O_p(T^{-\frac{3}{2}}). \end{aligned}$$

Thus, by lemma A.3, $E\left\{e'_p(\hat{A}_p^h - A_p^h)W_T v_T\right\} = O\left[E\left\{\|(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})'W_T v_T\|\right\}\right] = o(T^{-1})$. Let us denote $B'_{h,1} = e'_p \sum_{j=0}^{h-1} A_p^j(\hat{A}_p - A_p)A_p^{h-1-j}$. Then, by Hölders' inequality, $E\{e'_p(\hat{A}_p^h - A_p^h)W_T W'_T(\hat{A}_p^h - A_p^h)'e_p\} = E(B'_{h,1}W_T W'_T B_{h,1}) + O\left(T^{-\frac{3}{2}}\right)$. Applying theorem 4 and the result that the effect in the PMSE of the dependency between $\hat{\boldsymbol{\phi}}_{|p}$ and W_T is $O(T^{-\frac{3}{2}})$ (Kunitomo & Yamamoto, 1985), it follows that

$$E(B'_{h,1}W_T W'_T B_{h,1}) = (\sigma^2/T)\sum_{j=0}^{h-1}\sum_{k=0}^{j-1}(e'_p A_p^j e_p)(e'_p A_p^k e_p) \times \mathrm{tr}(A_p^{h-1-j}\Gamma_w {A'_p}^{h-1-k}\Gamma_w^{-1}) + o(T^{-1}),$$

and the proof of (5.6) is completed. Similarly, by the same arguments, expression (5.7) can be obtained. $\square$

D Proofs of section 5:

Proof of lemma 1: Let us decompose Y_t as $Y_t = (\tilde{Y}'_t,\ 0)' + \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu, \mu, ..., \mu, 1)'$. Since $\alpha = \mu(1 - \sum_{i=1}^{p+1}\varphi_i)$, it can be shown that $A_\alpha^i \boldsymbol{\mu} {A'_\alpha}^j = \bar{\boldsymbol{\mu}}$, where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}\boldsymbol{\mu}'$. Then $A_\alpha^i \Gamma_y {A'_\alpha}^j = A_\alpha^i \Gamma_{\tilde{y}}^* {A'_\alpha}^j + \bar{\boldsymbol{\mu}}$, where $\Gamma_{\tilde{y}}^*$ is a $(p+2)\times(p+2)$ matrix with $\Gamma_{\tilde{y}}$ in the first $(p+1)\times(p+1)$ submatrix and zero elsewhere. Also, the covariance matrix Γ_y has the following block structure

$$\Gamma_y = \begin{pmatrix} \Gamma_o & \boldsymbol{\mu}_o \\ \boldsymbol{\mu}'_o & 1 \end{pmatrix},$$

where $\Gamma_o = E(Y_{ot}Y'_{ot})$, with $Y_{ot} = (y_t, y_{t-1}, ..., y_{t-p})'$ and $\boldsymbol{\mu}_o = E(Y_{ot})$. Using the properties of the inverses of block matrices, we can partition Γ_y^{-1} as

$$\Gamma_y^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where $B_{11} = (\Gamma_o - \boldsymbol{\mu}_o\boldsymbol{\mu}'_o)^{-1} = \Gamma_{\tilde{y}}^{-1}$. Then it is verified that $\mathrm{tr}(A^i_\alpha \Gamma^*_{\tilde{y}} A'^{\,j}_\alpha \Gamma_y^{-1}) = \mathrm{tr}(A^i_o \Gamma_{\tilde{y}} A'^j_o \Gamma_{\tilde{y}}^{-1})$. Hence, $\mathrm{tr}(A^i_\alpha \Gamma_y A'^j_\alpha \Gamma_y^{-1}) = \mathrm{tr}(A^i_o \Gamma_{\tilde{y}} A'^j_o \Gamma_{\tilde{y}}^{-1}) + \mathrm{tr}(\boldsymbol{\mu}\Gamma_y^{-1})$. Given that $\mathrm{tr}(\bar{\boldsymbol{\mu}}\Gamma_y^{-1}) = \boldsymbol{\mu}'\Gamma_y^{-1}\boldsymbol{\mu}$, and applying a result of Searle (1984, pag. 258), it can be seen that $\boldsymbol{\mu}'\Gamma_y^{-1}\boldsymbol{\mu} = 1 - \mid \Gamma_y - \boldsymbol{\mu}\boldsymbol{\mu}' \mid / \mid \Gamma_y \mid = 1$, since the last column and row of $\Gamma_y - \boldsymbol{\mu}\boldsymbol{\mu}'$ are zero and Γ_y is invertible. □

Proof of lemma 2: Let C be the following nonsingular matrix

$$C = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\rho \\ 1 & -\phi_1 & -\phi_2 & \cdots & -\phi_{p-1} & -\phi_p \end{pmatrix}.$$

Then

$$D = CA_oC^{-1} = \begin{pmatrix} A_p & 0 \\ 0 & \rho \end{pmatrix}.$$

Let λ_k be an eigenvalue of the matrix $Q = \Gamma_{\tilde{y}}^{-1} A^i_o \Gamma_{\tilde{y}} A'^{\,j}_o$. Then

$$|D^i \Gamma_C D'^j - \lambda \Gamma_C| = 0, \tag{D.1}$$

where $\Gamma_C = C\Gamma_{\tilde{y}}C'$. This matrix Γ_C can be considered as the covariance matrix of the transformed series $Z_t = CY_t$, where $Z_t = (z_{1,t}, z_{1,t-1}, ..., z_{1,t-p+1}, z_{2,t})'$ and

$$Z_t = DZ_{t-1} + a_t c_{p+1}. \tag{D.2}$$

Therefore, the first $p \times p$ submatrix of Γ_C is the covariance matrix of a process $z_{1,t}$ following the coefficient matrix A_p and noise a_t; namely, the matrix $\Gamma_{w|p}$. Denoting by V_{12}, V_{21}, and V_{22} the remaining submatrices of this partitioning, we can rewrite (D.1) as

$$\begin{vmatrix} (A^i_p \Gamma_{w|p} A'^{\,j}_p - \lambda \Gamma_{w|p}) & (A^i_p V_{12} \rho^j - \lambda V_{12}) V_{22}^{-\frac{1}{2}} \\ (\rho^i V_{21} A'^{\,j}_p - \lambda V_{21}) V_{22}^{-\frac{1}{2}} & (\rho^{i+j} - \lambda) \end{vmatrix} = 0.$$

From (D.2), the term V_{22} is the variance of an AR(1) process with coefficient ρ. Therefore $V_{22}^{-1} = O(1-\rho)$. Hence, using the rule to evaluate the determinant of a partitioned matrix (see, for instance, Searle, 1984)

$$|Q - \lambda I| = \left| A^i_p \Gamma_{w|p} A'^{\,j}_p - \lambda \Gamma_{w|p} \right| \left\{ \rho^{i+j} + O(1-\rho) - \lambda \right\} = 0.$$

Since the trace of a matrix equals the sum of its eigenvalues, the lemma follows. □

References

Bhansali, R. J. (1981). "Effects of Not Knowing the Order of an Autoregressive Process on the Mean Squared Error of Prediction– I," *Journal of the American Statistical Association,* 76, 588–597.

Campbell, J. Y. & Perron (1991). "Pitfalls and Opportunities: What Macroeconomists should Know About Unit Roots ," *NBER Macroeconomics Annual*, 141–201.

Chan, N. H. & Wei, C. Z. (1987). "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics,* 15, 1050–1063.

Dickey, D. A., and Fuller, W. A. (1979). "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of the American Statistical Association,* 74, 427–431.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: John Wiley.

Fuller, W. A. & Hasza, D. P. (1981). "Properties of Predictors for Autoregressive Time Series," *Journal of the American Statistical Association,* 76, 155–161.

Granger, C.W.J. & Joyeux, R. (1980). "An Introduction to Long-range Time Series Models and Fractional Differencing," *Journal of Time Series Analysis,* 1, 15–30.

Harvey, A. C. (1981). "Finite Sample Prediction and Overdifferencing," *Journal of Time Series Analysis,* 2, 221–232.

Kunitomo, N. and Yamamoto, T. (1985). "Properties of Predictors in Misspecified Autoregressive Time Series Models," *Journal of the American Statistical Association,* 80, 941–950.

Ledolter, J. and Abraham, B.(1981). "Parsimony and Its Importance in Time Series Forecasting," *Technometrics,* 23, 411–414.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis.* Berlin: Springer-Verlag.

Phillips, P. C. B. (1987). "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika,* 74, 535–547.

Plosser, C. I. and Schwert, G. W.(1977). "Estimation of a Non-invertible Moving Average Process: The Case of Overdifferencing," *Journal of Econometrics,* 6, 199–224.

Plosser, C. I. and Schwert, G. W. (1978). "Money, Income and Sunspots: Measuring Economic Relationships and the Effects of Differencing," *Journal of Monetary Economics* **4** 637–660.

Shaman, P. and Stine, R. (1988). "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association,* 83, 842–848.

Searle, S. R. (1984). *Matrix Algebra Useful for Statistics.* New York. Wiley.

Stock, J. H. (1996). "VAR, Error Correction and Pretest Forecast at Long Horizons ," *Oxford Bulletin of Economics and Statistics* **58**, 685–701.