

# Testing long-horizon predictive ability with high persistence, and the Meese-Rogoff puzzle

Barbara Rossi\*

First draft: May, 2000      This draft: February, 2003

## Abstract

A well-known puzzle in the international finance literature is that a random walk predicts exchange rates better than economic models (Meese and Rogoff, 1983a,b and 1988). This paper offers a potential explanation for this finding. When exchange rates and fundamentals are highly persistent, long-horizon forecasts of economic models are asymptotically biased by the estimation error in the parameter that measures the persistence. When this bias outweighs the benefits from exploiting economic information, the random walk model will forecast better. This happens even if the economic model is the true data generating process. The reason is that a random walk model imposes a unit root, rather than estimates it. The paper thus proposes a test for equal predictive ability in the presence of highly persistent variables. When applied to the Meese-Rogoff exercise, this test shows that the poor forecasting ability of economic models *does not* imply that the models are *not* a good description of the data.

*Keywords: Predictive ability; random walk; long-horizon forecasting; roots close to unity.*

*JEL classification: F300; F400.*

*Acknowledgments.* This paper originated in discussions with Mark Watson, to whom I am sincerely grateful for his help, encouragement and teaching. I am also grateful to Kenneth Rogoff and Bo Honore' for many helpful suggestions along the way, Moshe Buchinsky for a code to implement quantile regression, Alessandro Tarozzi, Bobray Bordelon, Valentina Corradi, Gregory Chow, Lutz Kilian, Enrique Mendoza, Douglas Miller, Elena Pesavento, Ernst Schaumburg, Christopher Sims and seminar participants at Bocconi, CIRANO, Iowa State, JHU, Princeton, Rutgers, UCLA, UCSD, Vanderbilt and Wisconsin-Madison and my colleagues at Duke for comments. Financial support from the Mellon Sawyer Seminar and the IFS Summer Research, Princeton University, is gratefully acknowledged. All mistakes are mine.

\*Department of Economics, Duke University, Durham NC27708 USA.

E-mail: brossi@econ.duke.edu. Tel.: (919) 660 1801. Fax: (919) 684 8974.

# 1. Introduction

In an influential series of papers, Meese and Rogoff (1983a,b and 1988) showed that the economic models' forecasts of future nominal and real exchange rates were worse than those of a simple random walk. The result was striking, as the random walk model does not use any information on “fundamentals” (that is, economic variables other than the lagged exchange rate) and does not have an economic interpretation.<sup>1</sup> Even more surprisingly, the superiority of the random walk held for conditional out-of-sample forecasts as well, that is for forecasts that use realized (as opposed to forecasted) values of the fundamentals. Hence, the failure of the economic models could not be attributed to the poor ability of the models to predict variables other than the exchange rate, but only to the poor predictive ability of the fundamentals in forecasting the exchange rate. Subsequent research analyzed the robustness of these findings to different samples, model specifications and explanatory variables. Although the overall empirical evidence is mixed, the Meese-Rogoff finding of the poor forecasting ability of economic models relative to the random walk has never been convincingly overturned.

The first contribution of this paper is to show that the Meese-Rogoff puzzle can be resolved if the standard errors of tests for predictive ability appropriately take into account the presence of highly persistent and not-exactly cointegrated variables. The paper shows that, otherwise, one would incorrectly conclude that the random walk forecasts better than the economic model even when the economic model is the true data generating process. The analysis begins by reconsidering the Meese-Rogoff puzzle and reproducing their results. We find that the mean square forecast error (hereafter MSE) of economic models is always bigger than that of the random walk, and that the performance of the model relative to the random walk worsens as the horizon of prediction increases. Next, we proceed to examine the time series properties of the data. These series are highly persistent but not exactly cointegrated. This leads to parameter estimation error in the estimate of the persistence at long horizons that introduces an asymptotic *bias* in the estimation of the model.<sup>2</sup> The reason is that the closer the root is to unity, the more the parameter estimation error plagues the economic models' estimation (which rely on inconsistent estimates of the persistence) relative to random walks (which *impose*, rather than *estimate*, a unit root). Thus, the economic models forecast poorly, worse than the random walk, and the MSE difference between the model and the random walk has a median

---

<sup>1</sup>Some authors (see Roll, 1979) argue that a random walk model is a sensible model because real and nominal exchange rate changes, like asset prices, should not be predictable if foreign exchange markets are efficient. However, as Froot and Rogoff (1995) clarify, this analogy is inappropriate, since real exchange rates are not traded assets. Hence, they are not market variables, whose price is subject to arbitrage conditions. Nominal exchange rates are market variables, but there is no reason to expect them to be random walks in the presence of nominal interest rate differentials or risk premia.

<sup>2</sup>The explanation is kept at an intuitive level in this introduction. We will be more specific later on.

positive asymptotic bias. Moreover, as the horizon of forecast increases, the distribution of the MSE difference spreads out to take into account the additional future uncertainty. This implies that the forecasts of the model appear to worsen with the forecast horizon.

The second contribution of this paper is to propose a *test* for equal predictive ability when variables are highly persistent but not exactly cointegrated. Existing tests assume stationarity and ergodicity, so that a Central Limit Theorem can be applied, as in Diebold and Mariano (1995) and West (1996). However, the Central Limit Theorem does not provide an accurate approximation in small samples in the presence of highly persistent variables. Subsequent research by Corradi, Swanson and Olivetti (2001) focused on the case of exact cointegration whereas Berkowitz and Giorgianni (2001) analyzed the case of exact unit roots and no cointegration. This paper, instead, uses the more general local-to-unity asymptotic theory (as in Cavanagh, Elliott and Stock (1995)) to propose a test for equal predictive ability that is robust to the presence of high persistence.

When this test is applied to the Meese and Rogoff problem, we find that, even if the random walk forecasts better, the improvements are not statistically significant at all horizons and for the different currencies considered.

The finding that, in the presence of high persistence, forecasts of a random walk may perform better than those of a univariate autoregression at long horizons is not new in the literature. Stock (1996) provides convincing Monte Carlo simulations for the univariate case and theoretical results for cointegrated models, and Diebold and Kilian (2000) study the usefulness of unit-root tests as diagnostic tools for selecting forecasting models. Phillips (1998) and Kemp (1999) show that the distribution of forecast errors in the presence of high persistence is non-normal. However, to date, there are no general results on the implications of high persistence for tests of predictive ability.

The paper is organized as follows. The second section provides a simplified example to highlight the main insight of this paper and describes the data generating process and its assumptions. Section 3 derives the asymptotic distribution of tests for predictive ability in the presence of high persistence, section 4 discusses how to perform tests in this framework, and section 5 shows some simple Monte Carlo simulations. Section 6 contains the empirical evidence on the Meese and Rogoff puzzle and section 7 concludes.

## 2. The model: motivation and assumptions

The following simplified example highlights the main idea in this paper. Consider an economic model that describes a relationship between the nominal exchange rate,  $y_{1t}$ , and a (strictly exogenous) economic explanatory variable, or *fundamental* determinant,  $y_{2t}$ :

$$y_{1t} = \beta y_{2t} + \eta_t \tag{1}$$

The researcher, as in Meese and Rogoff (1983a,b and 1988), is interested in forecasting the future value of the first variable  $h$  periods ahead in the future,  $y_{1t+h}$ , conditional on knowing the future value of the second variable,  $y_{2t+h}$ . If the error sequence  $\{\eta_{t+i}\}_{i=1}^h$  were unpredictable, then the best forecast would be  $\beta y_{2t+h}$ . However, in the data the residuals are highly serially correlated:

$$\eta_t = \rho\eta_{t-1} + \epsilon_t \quad \rho \simeq 1 \quad (2)$$

In principle, forecasts can be improved by exploiting the information contained in the serial correlation. Since the best forecast of the residual  $h$  periods ahead is  $\rho^h\eta_t$ , and since  $\eta_t = y_{1t} - \beta y_{2t}$ , the best forecast based upon the economic model, in population, call it  $\widehat{y}_{1t+h}^m$ , would be  $\widehat{y}_{1t+h}^m = \beta y_{2t+h} + \rho^h\eta_t$ . Thus, the forecast error of the model is:

$$y_{1t+h} - \widehat{y}_{1t+h}^m = (1 - \rho^h)\eta_t + (\eta_{t+h} - \eta_t) \quad (3)$$

As an alternative model, consider the random walk, which assumes that the change in the exchange rate is unpredictable. The random walk forecast (call it  $\widehat{y}_{1t+h}^{rw}$ ) does not use any information on economic variables other than the lagged exchange rate:  $\widehat{y}_{1t+h}^{rw} = y_{1t}$ . Thus, conditional on the model being true, the forecast error of the random walk is:

$$y_{1t+h} - \widehat{y}_{1t+h}^{rw} = \beta(y_{2,t+h} - y_{2,t}) + (\eta_{t+h} - \eta_t) \quad (4)$$

Meese and Rogoff found that the random walk forecast errors (4) were “smaller” than those of the model (3). They concluded that this result casted doubts on the ability of existing exchange rate models to explain the exchange rate fluctuations.

However, when the variables are highly persistent and not exactly cointegrated (as is the case for exchange rates and their fundamentals – see Froot and Rogoff (1995)), the random walk model may indeed forecast better than the economic model even if the economic model is the true data generating process. The reason is that the economic model relies on an estimate of the persistence at long horizons,  $\rho^h$ , for which usual (pointwise) asymptotic theory may provide poor approximations in small samples. In other words, it cannot be measured with sufficient precision to obtain asymptotically (median) *unbiased* forecasts.<sup>3</sup> As a consequence, the attempt to improve the forecasts by taking into account the serial correlation in the data will result in an asymptotic “bias” in the long horizon forecasts of the model. Hence, the difference in the forecasting ability of the two models will reflect not only the possible bias in the forecast of the random walk (which “forgets” the existence

---

<sup>3</sup>To save space and for lack of better words, in this paper we will refer to this problem as a “bias”, but it is important to keep in mind that this problem does not disappear asymptotically. It results in prediction intervals with asymptotically incorrect coverage rates. For example, in the simple univariate random walk case, Phillips (1998) reports that there is an appreciable probability (0.68) that the estimated prediction error variance is less than the actual prediction error variance of the optimal predictor.

of a relationship between the exchange rate and the fundamental,  $\beta(y_{2,t+h} - y_{2,t})$  in (4)), but also the possible “bias” in the forecast of the model (which relies on the estimate of  $\rho^h$  in (3)). Which model will end up forecasting better will depend upon the relative importance of the two “biases”.

Furthermore, note that the two forecast errors are functions of highly persistent variables,  $\eta_t$  and  $y_{2t}$ . It follows that both forecast errors will be highly persistent as well, and their squared average will be increasing with the forecast horizon. Hence, in situations in which the bias of the model is sufficiently big relative to that of the random walk, it will appear as if the model’s performance is worsening as the forecast horizon increases. While this example is admittedly unrealistic, the rest of the paper will show that its main insights do carry over to more general situations.

### The model

Let the data generating process (hereafter DGP) be:

$$\begin{aligned} (I - \Phi L) w_t &= u_t \\ w_t &\equiv B y_t - \bar{y} - D t \end{aligned} \tag{5}$$

where  $B$  is an  $(n+1) \times (n+1)$  identity matrix except for the first row, which is equal to  $(1, \beta)'$ , where  $\beta$  is a  $n \times 1$  vector of parameters.  $y_t$  is a  $(n+1) \times 1$  vector of variables partitioned as  $y_t = (y_{1t}, y'_{2t})'$ , where  $y_{1t}$  is a scalar and  $y_{2t}$  is a  $(n \times 1)$  vector (and all matrices and vectors may be partitioned accordingly).  $\bar{y}$  is a  $((n+1) \times 1)$  vector of constants,  $L$  is the lag operator,  $u_t$  is a  $((n+1) \times 1)$  stationary and ergodic moving average sequence of finite order  $q$ :<sup>4</sup>

$$u_t = \Theta(L)\epsilon_t \tag{6}$$

and  $\epsilon_t$  is a martingale difference sequence that satisfies assumption (A) below. The time series may have a time trend. Thus, following Stock and Watson (1996), we allow  $y_t$  to have a small time trend component, so small that testing whether it is absent would fail to reject the null hypothesis with positive probability asymptotically. That is,  $D = (d_1, d'_2)'$ , where  $d_1$  is a scalar and  $d_2$  is a  $n \times 1$  vector such that  $D = T^{-1/2}\tilde{d}$  and  $\tilde{d} = (\tilde{d}_1, \tilde{d}'_2)'$  is a vector of constants.  $\Phi$  is a diagonal matrix with elements  $(\phi_1, \phi'_2)'$  on the main diagonal, where  $\phi_1$  is a scalar and  $\phi_2$  is a  $n \times 1$  vector. Finally, let  $\Psi$  denote the vector of all parameters.

In order to obtain better asymptotic approximations to the statistic of interest in the presence of high persistence, we adopt the following multivariate local-to-unity device (as in Stock (1991, 1996) and Phillips (1987)):

$$\Phi = I - \frac{1}{T}\mathbf{C} \tag{7}$$

---

<sup>4</sup>The ARMA representation for  $u_t$  is a flexible and convenient parametrization of the process. However, the main results of the paper can be generalized to allow  $u_t$  to be a generic strong mixing process (with sufficient restrictions on the serial dependence, as in Phillips (1988)).

where  $\mathbf{C}$  is a  $(n+1) \times (n+1)$  diagonal matrix with the vector  $(c_1, \mathbf{c}'_2)'$  on the main diagonal,  $\mathbf{c}'_2 = (c_{2,1}, c_{2,2}, \dots, c_{2,n})$ ,  $T$  is the sample size and  $I$  is the  $(n+1) \times (n+1)$  identity matrix. The process described by (5) has exact unit roots in case  $c_1 = 0, \mathbf{c}_2 = \mathbf{0}_{n \times 1}$ , where  $\mathbf{0}_{n \times 1}$  denotes a  $(n \times 1)$  vector of zeros; it is highly persistent in case they are both small; it is cointegrated when  $c_1 \rightarrow \infty, \mathbf{c}_2 = \mathbf{0}_{n \times 1}$ ; it is quasi-cointegrated when  $c_1 \rightarrow \infty, \mathbf{c}_2 \approx \mathbf{0}_{n \times 1}$ ; and it is stationary when  $c_1 \rightarrow \infty, c_{2,j} \rightarrow \infty \forall j$ . The diagonality of  $\Phi$  rules out processes integrated of order higher than one.<sup>5</sup> The fact that  $\Phi \simeq \mathbf{e}^{-\mathbf{C}/T}$ , where  $\mathbf{e}^{-\mathbf{C}/T}$  is a diagonal matrix with  $(e^{-c_1/T}, e^{-c_{2,1}/T}, \dots, e^{-c_{2,n}/T})'$  on the main diagonal, will be used (see Stock (1996)).

### Assumptions

We assume that (5) satisfies the following set of assumptions.

*Assumption A.*

(a)  $\epsilon_t$  is a martingale difference sequence:  $E(\epsilon_t | \epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_1) = 0$ .

(b)  $E(\epsilon_t \epsilon'_t | \epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_1) = \Sigma$ .

(c) Fourth moments are finite:  $\max_i E(\epsilon_{it}^4) < \infty$

(d)  $\Theta(L) = \sum_{i=0}^{\infty} \Theta_i L^i$  is a matrix polynomial in the lag operator with

$$\sum_{i=0}^{\infty} i |\Theta_i| < \infty, \Theta_0 = I \text{ and } \Lambda \equiv \Theta(1)\Sigma^{1/2} \text{ is invertible}$$

The researcher is interested in forecasting  $y_1$  at some horizon  $h$ , where  $h$  is called the horizon of prediction. We assume that the researcher consistently estimates the parameters of the model. In fact, the point of this paper is that, *even when the parameters are consistently estimated*, a high degree of persistence might cause the economic model's forecasts to be quite poor relative to those of a simple random walk model. The rates of convergence are described by the following assumption. Estimates obtained by using reduced-form VARs as well as Cochrane-Orcutt satisfy this assumption. Note that rolling, recursive or split-sample estimation methods would not affect the rates of convergence.<sup>6</sup>

---

<sup>5</sup>Note that the diagonality assumption rules out some models, for example Augmented Distributed Lags models in levels without restrictions across the parameters. However, the assumption is made in order to match the empirical characteristics of the data: the economic variables considered in this paper are highly persistent, but not integrated of order two.

<sup>6</sup>The proof for the VAR is in the appendix whereas the proof for the Cochrane-Orcutt is in an appendix available upon request. Depending on the estimation method, these assumptions may be stronger than necessary: for example, if one uses Cochrane-Orcutt then only parameters in the first equation need to be estimated so the assumptions on the additional parameters become redundant. Also, reduced-form VARs do not actually estimate  $\beta$ , but rates of convergence would be the same for their counterpart-“starred” variables defined in the appendix. E.g. the rate for  $\Phi^* \equiv B^{-1}\Phi B$  is that for  $\Phi$ . Finally, rolling, recursive, and split-sample estimates would result in different  $o_p(1)$  terms, but the qualitative results and the conclusions of the paper would be the same.

*Assumption B*

Rewrite the DGP as  $\Theta(L)^{-1}(I - \Phi L)By_t = \mu_0 + \mu_1 t + \epsilon_t$ . Then:

- (a)  $\hat{\beta} - \beta = O_p(\sqrt{T})$
- (b)  $\hat{\mu}_0 - \mu_0 = O_p(\sqrt{T})$
- (c)  $\hat{\Theta}(L) - \Theta(L) = O_p(\sqrt{T})$
- (d)  $\hat{\Phi} - \Phi = O_p(T)$
- (e)  $\hat{\mu}_1 - \mu_1 = O_p(T^{3/2})$

where “ $\hat{\psi}$ ” denotes the estimate of the parameter  $\psi$ .

Finally, this paper assumes that the number of predictions ( $P$ ) is a fixed fraction of the sample size and that the horizon of prediction is a fixed fraction of the sample size too:

*Assumption C:*

$$\frac{P}{T} \rightarrow \pi \tag{8}$$

$$\frac{h}{T} \rightarrow \delta \tag{9}$$

Condition (8) is standard (see West (1996)), whereas condition (9) is introduced here in order to obtain better approximations to the asymptotic distribution of the test statistics when the horizon of prediction is big relative to the sample size (see Richardson and Stock (1989), Stock (1996) and Phillips (1998)). This device turns out to be particularly useful in analysis of long horizon exchange rates, as databases for major currencies during the recent floating period hardly exceed one hundred observations for quarterly data. As a special case, when  $\delta = 0$  then the asymptotic distribution is the same as in Diebold and Mariano (1995) and West (1996).

The following lemma describes the asymptotic distribution of the DGP under assumption (A):

*Lemma 1. Asymptotic distribution.*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} \epsilon_t \Rightarrow \Sigma^{1/2} W^*(r); \quad \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} u_t \Rightarrow W(r) = \Lambda W^*(r); \tag{10}$$

$$\frac{1}{\sqrt{T}} w_{[Tr]} \Rightarrow \Lambda J_c(r); \quad dJ_c(r) = -\mathbf{C}J_c(r)dr + dW(r) \tag{11}$$

$$\frac{1}{\sqrt{T}} y_{[Tr]} \Rightarrow \tilde{J}_c(r); \quad \tilde{J}_c(r) = B^{-1}(\delta r + \Lambda J_c(r)); \tag{12}$$

where  $\Rightarrow$  denotes weak convergence,  $W^*(r)$  is a standardized Brownian motion vector process,  $J_c(\cdot)$  is a standardized Ornstein-Uhlenbeck vector process and  $[.]$  is the greatest lesser integer function. Proofs are provided in Phillips (1987) for the univariate case with  $\tilde{d} = 0$  and Stock and Watson (1996) for the more general case where  $\tilde{d} \neq 0$ .

### 3. Distribution of test statistics for predictive ability

The existing literature commonly measures the predictive ability of a model by the *Mean Squared forecast Error* (hereafter MSE):

$$MSE_h^\kappa \equiv \frac{1}{P} \sum_{t=R}^{T-h} (y_{1,t+h} - \widehat{y}_{1,t+h}^\kappa)^2 \quad (13)$$

where notation is as follows:  $y_{1,t+h}$  is the realized value of variable  $y_1$  (the variable that one is interested in forecasting) at time  $t+h$ ; superscript “ $\kappa$ ” refers to the model used for forecasting ( $\kappa = m$  for the economic model and  $\kappa = rw$  for the random walk);  $\widehat{y}_{1,t+h}^\kappa$  is the forecast of that variable made at time  $t$  according to model  $\kappa$ ;  $y_{1,t+h} - \widehat{y}_{1,t+h}^\kappa$  is the forecast error;  $P$  is the total number of predictions, which are based on a parameter vector estimated using data at most up to  $t$ ;<sup>7</sup>  $R$  is the size of the sample used for the first estimation; the size of the available sample is  $R + P - 1 + h = T$ .<sup>8</sup> It is then natural to compare the predictive ability of two competing models, say the economic model “ $m$ ” and the random walk “ $rw$ ”, by the *Difference in the MSEs* of their forecasts, hereafter denoted by:

$$dm \equiv MSE_h^m - MSE_h^{rw} \quad (14)$$

We are interested in testing equal predictive ability, i.e. to test the null hypothesis  $E(dm) = 0$  versus the alternative  $E(dm) \neq 0$ .

If the variables are strictly stationary and ergodic, the Central Limit Theorem can be applied to (13) so that  $\sqrt{P}[dm - E(dm)] \Rightarrow N(0, V)$ , where  $V$  is the asymptotic variance of the squared forecast error difference. This is the tests proposed by West (1996) and Diebold and Mariano (1995). However, as the standard Central Limit Theorem does not provide accurate approximations in small samples in the presence of highly persistent variables, standard tests for predictive ability that rely on that may have size distortions. Thus, we next derive the asymptotic distribution of (14); the size distortions of standard tests for predictive ability will be evaluated in a Monte Carlo example.

By using the Wiener-Kolmogorov prediction formula (see the appendix), the DGP (5) can be approximated as:

$$w_{t+h} = \underbrace{\sum_{i=0}^{h-1} \Phi^i \Theta(I) \epsilon_{t+h-i}}_{e_{t+h}(\Psi|\mathfrak{S}_t)} + \underbrace{\Phi^h w_t}_{\widehat{w}_{t+h}(\Psi|\mathfrak{S}_t)} + o_p(T^{1/2}) \quad (15)$$

which shows that the value of  $w_{t+h}$  can be decomposed in the sum of a forecast error (denoted by  $e_{t+h}(\Psi|\mathfrak{S}_t)$ ) and a predictable component at time  $t$  (denoted by  $\widehat{w}_{t+h}(\Psi|\mathfrak{S}_t)$ ), where  $\mathfrak{S}_t = \{y_s\}_{s=1}^t$  denotes the information set at time  $t$ .<sup>9</sup> Different forecasting methods may rely on different

<sup>7</sup>Rolling, recursive and split-sample estimation are special cases.

<sup>8</sup>Note that this and (8) imply that  $\frac{R}{T} \rightarrow 1 - \pi - \delta$ , so  $R$  is a fixed fraction of the sample size as well.

<sup>9</sup>These components depend on both the data as well as the parameters  $\Psi$ . For brevity, the dependence on the data will be dropped.

information sets  $\mathfrak{S}$  (e.g. ex-ante versus forecasts conditional on realized values of the regressors) and, thus have, in general, different forecast errors and predictable components. We will use the Selection matrix  $S$  to obtain algebraic expressions of forecast errors and predictable components associated to different information sets. In particular, forecasts conditional on realized values of the regressors rely on the information set  $\mathfrak{S}_{t+h} \equiv \{y_{1s}\}_{s=1}^t \cup \{y_{2s}\}_{s=1}^{t+h}$ , which corresponds to  $S = B$ ,<sup>10</sup> whereas ex-ante forecasts rely on the information set  $\mathfrak{S}_t \equiv \{y_s\}_{s=1}^t$ , which corresponds to  $S = I$ .

The forecasts of the model focus on the first component of (15),  $i_1' w_{t+h}$ , where  $i_1 = [1, \mathbf{0}_{1 \times n}]'$ . Thus, the forecasting error when using the model and the information set  $\mathfrak{S}$  and assuming that the parameter values are known, denoted by  $e_{t+h}^m(\Psi|\mathfrak{S}) \equiv i_1' e_{t+h}(\Psi|\mathfrak{S})$ , is:

$$\frac{e_{t+h}^m(\Psi|\mathfrak{S})}{\sqrt{h}} \equiv \frac{1}{\sqrt{h}} \sum_{i=0}^{h-1} i_1' S B^{-1} \Phi^i \Theta(I) \epsilon_{t+h-i} \quad (16)$$

It follows from Lemma 1 and the continuous mapping theorem (see also Stock (1996)) that:

$$\frac{1}{h} (e_{t+h}^m(\Psi|\mathfrak{S}))^2 \Rightarrow \omega_m^2(\tau) \quad (17)$$

where  $\omega_m(\tau) \equiv i_1' S B^{-1} \int_0^\delta e^{-\mathbf{C}s} dW(\tau + \delta - s) ds$  where  $\tau \equiv t/T$  and  $\mathbf{e}^{-\mathbf{C}s}$  is a diagonal matrix with  $(e^{-c_1 s}, e^{-c_{2,1} s}, \dots, e^{-c_{2,n} s})$  on the main diagonal.

To derive an expression for the forecast error of the random walk, recall from (5) that  $y_t = B^{-1} w_t + \bar{y} + Dt$ , so that  $y_{t+h} - y_t = B^{-1} (w_{t+h} - w_t) + Dh$ , and use (15) to write the forecast error of the random walk,  $e_{t+h}^{rw}(\Psi) \equiv i_1' (y_{t+h} - y_t)$ , as:

$$\frac{e_{t+h}^{rw}(\Psi)}{\sqrt{h}} \equiv h^{-1/2} i_1' [B^{-1} (w_{t+h} - w_t) + hD] \quad (18)$$

$$= h^{-1/2} i_1' [B^{-1} e_{t+h}(\Psi|\mathfrak{S}_t) + B^{-1} (\Phi^h - I) w_t + hD] \quad (19)$$

The MSE is the sum of the forecast error variance and the ‘‘bias’’ squared. In the expressions above, the first component can be interpreted as contributing to the variance and the second to the bias (it is known, conditional on information at time  $t$ ). Notice that the two components are independent, but highly persistent. The asymptotic distribution, conditional upon information at time  $t$ , then is:

$$\frac{1}{h} (e_{t+h}^{rw}(\Psi))^2 \Rightarrow (bias_{rw}(\tau) + \omega_{rw}(\tau))^2 \quad (20)$$

where  $h^{-1/2} i_1' [B^{-1} (\Phi^h - I) w_t] + (h/T)^{1/2} \tilde{d}_1 \Rightarrow i_1' [B^{-1} (e^{-\mathbf{C}\delta} - I) \Lambda J_c(\tau)] + \sqrt{\delta} \tilde{d}_1 \equiv bias_{rw}(\tau)$  and  $h^{-1/2} i_1' B^{-1} e_{t+h}(\Psi|\mathfrak{S}) = h^{-1/2} i_1' B^{-1} \sum_{i=0}^{h-1} \Phi^i \Theta(I) \epsilon_{t+h-i} \Rightarrow i_1' B^{-1} \int_0^\delta \mathbf{e}^{-\mathbf{C}s} dW(\tau + \delta - s) ds \equiv \omega_{rw}(\tau)$ .

So far the discussion abstracted from parameter estimation error. However, the model’s forecastable component,  $\hat{w}_{t+h}(\Psi|\mathfrak{S})$ , involves an estimate of  $\Phi^h$ , whereas the random walk imposes a

<sup>10</sup>When doing forecasts conditional on realized contemporaneous values of the regressors (by using, say, Cochrane-Orcutt), a strict exogeneity assumption is sufficient to ensure consistent estimation of the parameters. This implies that  $B$  and  $\Theta(L)$  are upper triangular.

unit root. It is well known (cfr. Stock (1996) and Phillips (1998)) that long horizon forecasts depend on the largest autoregressive roots of the system,  $\Phi$ , and that these largest roots cannot be measured with sufficient precision to obtain prediction intervals with asymptotically correct coverage rates over the possible values of the largest roots. To obtain better asymptotic approximations, we let  $\Phi^h = (I - \frac{1}{T}\mathbf{C})^h = (I - \frac{1}{T}\mathbf{C})^{T(h/T)}$  be approximated by  $\mathbf{e}^{-\mathbf{C}\delta}$  as  $\frac{h}{T} \rightarrow \delta$ . As all other parameters (apart from the deterministic components) can be consistently estimated, their effect is negligible and will be ignored. To evaluate the consequences of parameter estimation error, denote the estimate of the error of the model,  $e_{t+h}^m(\Psi|\mathfrak{S})$ , by  $e_{t+h}^m(\widehat{\Psi}|\mathfrak{S})$ . From assumption *B* it follows that:

$$e_{t+h}^m(\widehat{\Psi}|\mathfrak{S}) = e_{t+h}^m(\Psi|\mathfrak{S}) + i_1' SB^{-1} \left( \Phi^h - \widehat{\Phi}^h \right) w_t + \sqrt{T} \widehat{\zeta}_{pe}^{\det} + o_p(\sqrt{T}) \quad (21)$$

where  $\widehat{\zeta}_{pe}^{\det}$  denotes the parameter estimation error due to the estimation of the deterministic component at time  $t$  (see the appendix for details). From Lemma 1,  $i_1' T^{-1/2} w_t \Rightarrow i_1' \Lambda J_c(\tau)$ . Thus:

$$\frac{1}{\sqrt{h}} \left( e_{t+h}^m(\widehat{\Psi}|\mathfrak{S}) - e_{t+h}^m(\Psi|\mathfrak{S}) \right) \Rightarrow \frac{1}{\sqrt{\delta}} [i_1' SB^{-1} (\mathbf{e}^{-\mathbf{C}(\tau)\delta} - \mathbf{e}^{-\widetilde{\mathbf{C}}\delta}) \Lambda J_c(\tau) + \zeta_{pe}^{\det}(\tau)] \equiv bias_m(\tau) \quad (22)$$

where  $\mathbf{e}^{-\widetilde{\mathbf{C}}(\tau)\delta}$  is the (random) limiting distribution of the estimate of  $\Phi^h$  estimated at time  $[\tau T]$  and  $\zeta_{pe}^{\det}(\tau)$  is that of  $\widehat{\zeta}_{pe}^{\det}$ . Thus the estimated MSE of the model relative to the MSE of the model evaluated at the true parameter values has an asymptotic positive bias.

Note that if the model is estimated in first differences rather than in levels, then the unit roots are imposed rather than estimated. Thus, (22) may not be asymptotically relevant. In this case, the models will be nested, the null hypothesis one-sided and the asymptotic distribution will be different (see McCracken (1999) and Clark and McCracken (2001c)). Clark and McCracken (2001a) derive long horizon tests for predictive ability for nested models which apply to this situation (although without assumption (9)). In a related paper, Rossi (2002) analyzes the power of tests of predictive ability for models estimated in first differences and shows that the evidence in favor of the random walk is weakened. When the model is estimated in first differences, the empirical evidence points to parameter instability as an explanation for the forecasting failure of the economic models estimated by Meese and Rogoff.

Note also that in the predictive ability literature, whether  $P/R = 0$  or not is important about the possible contribution of parameter estimation error (as in the former case  $R \rightarrow \infty$  so the parameters can be treated as known). Here, even if  $R \rightarrow \infty$ ,  $\Phi^h$  cannot be treated as known (as we need convergence along the sequence of DGPs defined by the local to unity assumption) unless  $\delta = 0$ , in which case parameter estimation error in the roots will be asymptotically negligible.

The MSE is the average of the forecast errors over all predictions  $P$ , so the distribution of  $dm$  is a straightforward application of the continuous mapping theorem to the previous results:

**Proposition 1**<sup>11</sup>

Suppose that (5) and assumptions (A), (B) and (C) hold. Let  $\delta \neq 0$  and  $\pi \neq 0$ . Then:

$$\frac{1}{h}MSE_h^m \Rightarrow \int_{1-\delta-\pi}^{1-\delta} (bias_m(\tau) + \omega_m(\tau))^2 d\tau \quad (23)$$

$$\frac{1}{h}MSE_h^{rw} \Rightarrow \int_{1-\delta-\pi}^{1-\delta} (bias_{rw}(\tau) + \omega_{rw}(\tau))^2 d\tau \quad (24)$$

$$\frac{1}{h}dm \Rightarrow F_{\mathbf{C}, \Psi, \delta} \quad (25)$$

where  $\Rightarrow$  denotes weak convergence and the notation  $F_{\mathbf{C}, \Psi, \delta}$  emphasizes that the asymptotic distributions of the  $dm$  statistic is a function of  $\mathbf{C}, \Psi$  and  $\delta$ :

$$F_{\mathbf{C}, \Psi, \delta} \equiv \int_{1-\delta-\pi}^{1-\delta} \left\{ [\omega_m(\tau) + bias_m(\tau)]^2 - [\omega_{rw}(\tau) + bias_{rw}(\tau)]^2 \right\} d\tau \quad (26)$$

According to proposition 1, the asymptotic distribution of the statistics used to compare predictive ability is non-normal. Thus prediction intervals based on the central limit theorem will have actual coverage rates different from the nominal coverage rate. Note that the asymptotic distribution of the  $dm$  statistic (26) depends on two components. If  $bias_m(\tau) = 0 \forall \tau$  then  $\omega_m(\tau)^2 - [\omega_{rw}(\tau) + bias_{rw}(\tau)]^2$  would capture the dimension where the model has potentially an advantage over the random walk, since it can exploit information on the fundamentals cleaned from any parameter estimation error. However, estimation of the model introduces a first order bias,  $bias_m(\tau)$ , because of the inconsistent estimate of the unit root component at long horizons.

In general, these statistics are functionals of diffusion processes and depend on the unknown nuisance parameters  $c_1$  and  $\mathbf{c}_2$  in a complicated way. For extremes values of the nuisance parameters, such as strict stationarity or cointegration, the form of the distributions is known.

When the DGP is strictly stationary,<sup>12</sup> that is  $\phi_1, \phi_2, \dots, \phi_n$  are all in absolute value less than one, then  $e_{t+h}^m(\Psi|\mathfrak{F})$  defined in (16) is a stationary variable, which obeys a law of large numbers. Thus the results of Diebold and Mariano (1995) and West (1996) apply and, if the model is correctly specified, it forecasts are better (in population) than the random walk (this is the case in the Monte Carlo simulations shown in section 5). In the case of exact cointegration ( $\Phi_{11} \ll 1, \Phi_{22} = I$ ) then

<sup>11</sup>We do not report results on the fixed horizon asymptotics (where  $h$  is a fixed number, i.e.  $\delta = 0$ ) as the asymptotic distribution of the  $dm$  statistic in that case is the same as that of Diebold and Mariano (1995). Note that the two bias and variance components in (26) can, respectively, be correlated.

<sup>12</sup>We abstract from the existence of a time trend here; if a time trend exists and the random walk does not estimate it, the model will eventually perform better than the random walk at long horizons. We furthermore assume that, if the DGP includes a constant, the random walk is estimated with a drift.

$e_{t+h}^m(\Psi|\mathfrak{S}) = O_p(1)$  and  $e_{t+h}^{rw}(\Psi) = O_p(h^{1/2})$  so that the random walk will perform worse than the model, and the performance will be worsening with the horizon of prediction ( $\sqrt{P}dm \rightarrow -\infty$ ). This result is not inconsistent with the result obtained by Corradi, Swanson and Olivetti (2001): the models that these authors compared were both cointegrated, whereas in the framework of this paper one of them (the random walk) is not.

The analysis becomes more complicated in intermediate cases, since the asymptotic distribution of  $dm$  depends on the nuisance parameters and this complicates inference and testing. Section 5 will provide an illustrative Monte Carlo example.

#### 4. Testing equal predictive ability.

From Proposition 1, conventional critical values of tests for predictive ability are not valid in the presence of roots close to unity, and will depend on nuisance parameters ( $\mathbf{C}$ ) that cannot be consistently estimated. We construct valid critical values by using Bonferroni methods (see Cavanagh, Elliott and Stock (1995)).

Bonferroni critical values are obtained as follows: (i) in a first stage, construct a  $100(1 - \alpha_1)\%$  joint confidence set  $C_{\mathbf{C}}(\alpha_1)$  for the nuisance parameters; in a second stage, derive the acceptance region with size  $\alpha_2$  for  $dm$ ,  $C_{dm|\mathbf{C}}(\alpha_2)$ , as a function of the nuisance parameters under the hypothesis that the economic model is the true DGP; (iii) finally, an acceptance region for  $dm$  that does not depend on the nuisance parameters is the union (denoted by  $\cup$ ) of the second stage sets:

$$C_{dm}(\alpha) = \bigcup_{c \in C_{\mathbf{C}}(\alpha_1)} C_{dm|\mathbf{C}}(\alpha_2) \quad (27)$$

which, by construction, has size of at most  $100(\alpha_1 + \alpha_2)\%$ .

The confidence set  $C_{\mathbf{C}}(\alpha_1)$  is based on Stock and Watson's (1988) test statistic for common trends. The DGP, as in (5), is  $w_t = \Phi w_{t-1} + u_t$ , where the limiting variance of the process is  $\Lambda$ . It follows that:

$$\Phi_c \equiv \left( \frac{1}{T^2} \sum_{t=1}^T w_t w'_{t-1} - \frac{1}{T} M' \right) \left( \frac{1}{T^2} \sum_{t=1}^T w_{t-1} w'_{t-1} \right)^{-1} \quad (28)$$

where  $M = \sum_{j=1}^{\infty} E(u_{t-j} u'_t)$  can be consistently estimated by using Newey and West (1987).<sup>13</sup>

From lemma 1, Stock and Watson (1988) and a generalization of the results in Phillips (1987) to vector processes, it follows that:

$$T(\Phi_c - \Phi) \Rightarrow (\Lambda \psi'_k \Lambda') (\Lambda \Gamma_k \Lambda')^{-1} = \Lambda \psi'_k \Gamma_k^{-1} \Lambda^{-1} \quad (29)$$

---

<sup>13</sup>As in Stock and Watson (1988), the estimate of  $M$ ,  $\widehat{M}$ , is:  $\widehat{M} = \sum_{j=1}^{\infty} \left(1 - \frac{j}{q+1}\right) \widehat{V}'_j$ , where  $\widehat{V}_j = T^{-1} \sum_{t=j+1}^T \widehat{u}_t \widehat{u}'_{t-j}$  and  $q/T^{1/4} \rightarrow 0$ . Under assumptions (A) and (B), the estimate is consistent.

where  $\psi'_k = \int_0^1 J_c(r) dW_c(r)' dr$ ,  $\Gamma_k = \int_0^1 J_c(r) J_c(r)' dr$ . Let  $\text{eig}(T(\Phi_c - I))$  denote the (diagonal) eigenvalue matrix of  $T(\Phi_c - I)$ . It follows from (29) and the diagonality of  $\mathbf{C}$  that:

$$T\{\text{eig}(\Phi_c) - \mathbf{I}\} \Rightarrow \text{eig}(\psi'_k \Gamma_k^{-1}) - \mathbf{C} \quad (30)$$

which involves only functionals of standardized Ornstein-Uhlenbeck processes and can then be tabulated.<sup>14</sup> The confidence set  $C_{\mathbf{C}}(\alpha_1)$  is obtained by inverting (30) for the smallest real eigenvalue in absolute value (see Stock (1991)). That is, we simulate (30) as a function of  $(c_1, \mathbf{c}_2)$ . Then, given the value of the test statistic (30) estimated in the data,  $C_{\mathbf{C}}(\alpha_1)$  is the set  $(c_1, \mathbf{c}_2)$  such that the estimated test statistic belongs to the simulated quantiles.

The statistic for testing equal predictive ability,  $dm$ , depends on covariances and correlation parameters and thus it must be simulated conditionally on the estimated values of these parameters.

## 5. A small Monte Carlo illustrative example

A small Monte Carlo will be helpful in understanding the behavior of  $dm$ . For simplicity, let  $n = 1$  and  $\Theta(L)$  be diagonal, so that the existence of correlation among  $y_{1,t}$  and  $y_{2,t}$  relies exclusively on  $\beta$ . Thus, the DGP is simplified as:

$$\begin{aligned} y_{1t} &= \beta y_{2t} + u_{1t}; & y_{2t} &= u_{2t} \\ u_{1t} &= \rho_1 u_{1t-1} + \epsilon_{1t}, & \rho_1 &= 1 - c_1/T \\ u_{2t} &= \rho_2 u_{2t-1} + \epsilon_{2t}, & \rho_2 &= 1 - c_2/T \end{aligned} \quad (31)$$

and  $\epsilon_{1t}, \epsilon_{2t}$  are two uncorrelated and serially uncorrelated white noise processes, normally distributed with unit variance.

Figure 1 plots the distribution of the  $dm$  statistic for this DGP as a function of  $c_1$  and  $\mathbf{c}_2$  for  $\beta = -0.05$ . Figure 1 shows that the distribution of  $dm$  shifts to the right as  $c_1$  and  $c_2$  become smaller. Hence, for example, an observed value of the  $dm$  statistic equal to 0.2 would imply a rejection of the null hypothesis of equal predictive ability when the series are highly stationary (from table 1, critical values are (-0.4866; 0.0801) for  $c_1 = c_2 = 50$ ) in favor of the hypothesis that the model predicts better at 10% significance level. However, the same value of the statistic would accept the same hypothesis at the same confidence level when the series are highly persistent (for example,

<sup>14</sup>In fact, let  $(E, V)$  be the eigenvalue-eigenvector matrices of  $(\Phi_c - \Phi)$ . Then  $(\Phi_c - \Phi)V = EV$  so  $\Phi_c V = (E + I - \frac{1}{T}\mathbf{C})V$ . Let  $\lambda_j$  be the  $j$ -th eigenvalue of  $\psi'_k \Gamma_k^{-1}$ , which, by construction, is also the eigenvalue of  $T(\Phi_c - \Phi)$ . Thus, the  $j$ -th eigenvalue of  $\Phi_c$ , call it  $\lambda_{(j)}^{\Phi_c}$ , equals  $\lambda_j + 1 - \frac{1}{T}c_{(j)}$ , and  $T(\lambda_{(j)}^{\Phi_c} - 1) = T\lambda_j - c_{(j)}$ . We checked that the quantiles corresponding to exact unit roots ( $\mathbf{C}=\mathbf{0}$ ) were similar to those obtained by Stock and Watson (1988).

when  $c_1 = c_2 = 1$  the critical values become  $(-0.1087; 0.3206)$ . By using the former critical values, we would (wrongly) reject the null hypothesis.<sup>15</sup>

Insert Figure 1 and Table 1

Figure 2, instead, plots the distribution of the  $dm$  statistic for various values of  $\beta$  but for given values of  $c_1$  and  $c_2$ . Again, the distribution shifts to the right and becomes more asymmetric as  $\beta$  decreases. This behavior is not surprising, as  $\beta$  reflects the degree by which  $y_{1,t}$  and  $y_{2,t}$  are correlated: the less the series are correlated with each other, the smaller is the advantage of using  $y_{2,t}$  to predict  $y_{1,t}$ .

Finally, figure 3 shows that the  $dm$  statistic spreads out as the horizon of prediction increases and, as a consequence, the (median) positive asymptotic bias increases as the horizon increases.

Insert Figures 2 and 3

Overall, these simulations show that there is a trade-off between an economic model and the random walk concerning their ability to forecast out-of-sample. Although the economic model has the advantage of exploiting the information on (known) future values of the fundamentals, the same model suffers from the impossibility of consistently estimating the (high) persistence in the data. Empirical evidence that the random walk has a lower MSE than an economic model does not necessarily mean, then, that the model is ‘wrong’, but might just be caused by the interplay of a high persistence and a low correlation in the series.

## 6. The Meese and Rogoff puzzle revisited.

In this section, we apply the methods discussed so far in order to show that, by correctly taking into account sampling error, the forecasts of the random walk are not significantly better than those of existing economic models. Following Meese and Rogoff, the economic models considered in this paper are:

- the *Real Interest Rate Parity (RIRP)* condition, which states that bilateral *real* exchange rate fluctuations are driven by real interest rate differentials (Meese and Rogoff, 1988);
- the *Monetary* model, according to which bilateral *nominal* exchange rates are driven by money supplies, real incomes and short-term interest rates differentials (Meese and Rogoff, 1983a).

---

<sup>15</sup>The asymptotic distribution in the cointegration and quasi-cointegration cases is similar to the one in the stationary case and it is not reported. This result is due to Corradi, Swanson and Olivetti (2001).

Each economic model can be written as (5) by letting  $y_{1t}$  denote the logarithm of the exchange rate and  $y_{2t}$  denote the vector containing its economic explanatory variables (hereafter referred to as “fundamentals”). All variables, except interest rates, are in logarithm. The models are estimated in the same way as Meese and Rogoff did, so that the results are comparable to theirs.<sup>16</sup> Rolling forecasts of the exchange rate  $h$ -periods ahead are calculated out-of-sample and the predictive ability of the various models at an horizon of  $h$ -periods ahead is measured by the  $dm$  statistic (14). We consider two bilateral currencies versus the U.S. Dollar: the Deutsche Mark (hereafter D-Mark) and the Japanese Yen (Yen).<sup>17</sup> Data are quarterly, from 1973:3 to 1998:2, from Datastream.

Let’s consider first the empirical evidence on the Real Interest Rate Parity (RIRP) condition. Table 2 reports the relevant statistics. Note that  $dm$  is always *positive* at all horizons and *increasing* with the horizon. The same pattern emerges from visual inspection of figure 4, which compares the forecasts based on the RIRP with those of the random walk for the D-Mark data (results are qualitatively similar for the Yen). This confirms Meese and Rogoff (1988) findings.

Insert Table 2 and Figure 4

But are the forecasts of the model significantly worse than those of the random walk? We construct Bonferroni acceptance regions for equal predictive ability as described in this paper. First, we construct  $C_{\mathbf{C}}(\alpha_1)$ . The simulated values of the 1.25% and 98.75% quantiles of the minimum eigenvalue of (30) are a monotone function of the nuisance parameters  $(c_1, c_2)$ <sup>18</sup> and decrease as the

<sup>16</sup>The parameters are estimated in rolling samples by iterative Cochrane-Orcutt. In the iterative Cochrane-Orcutt, the step that involve estimation of the serial correlation are estimated by OLS whereas the steps that involve estimation of the relationship between the exchange rate and the fundamentals are estimated by efficient GMM. The instruments are the fundamentals and the exchange rate lagged one period. Note that, although we used the lagged exchange rate for estimation, we made sure that the latest lagged exchange rates did not feed back in the forecasting exercise, otherwise this feedback would give an unfair advantage to the economic model.

<sup>17</sup>The simulations of the  $dm$  test statistic are conditional upon the estimated parameters so they are performed specifically for each model and each currency. This is the reason why only two bilateral exchange rates are considered. However, these involve the US dollar, the D-Mark and the Japanese Yen are the currencies used for most transactions in the seventies and eighties. Furthermore, this paper focuses on methodological issues, so it did not seem worthwhile to consider additional currencies.

<sup>18</sup>The simulations were performed over a coarse grid of  $60 \times 60$  values for each dimension of  $\mathbf{C}$ , with values ranging from 1 to 30 with increments of one-half. The quantiles are obtained by using the quantile regression method of Koenker and Bassett. This method requires less simulation time than the alternative method of simulating the distribution of the test statistic for each value of  $(c_1, c_2)$  and retrieving the quantiles. A preliminary investigation compared both methods and the results were similar; if only, the Koenker and Bassett method was slightly biased at the boundaries so we simulated over a bigger grid than that of interest and we subsequently deleted the values at the boundaries. The code used for quantile regression was written by Moshe Buchinsky and provided by Bo Honore’. The regression includes a polynomial of second order and the values of both the dependent and the independent variables

values of the nuisance parameters increase, that is as the process approaches stationarity. Figure 5 shows  $C_{\mathbf{C}}(\alpha_1)$ , which corresponds to the area between the origin and the frontier plotted in the picture (the frontier in figure 5 corresponds to the .9875 quantile, as the .0125 quantile is never binding for both currencies). Then we construct  $C_{dm|\mathbf{C}}(\alpha_2)$ . Selected values of the  $dm$  statistic for values of the nuisance parameters inside  $C_{\mathbf{C}}(\alpha_1)$  are reported in table 3.

Insert Figure 5 and Table 3

Because of the monotonicity of these functions, it follows that one can build a Bonferroni acceptance region for, say, one-step ahead forecasts of size equal to at least 5% by taking respectively the minimum and maximum values of the third and fourth columns (the percentiles of the  $dm$  statistic) that appear in table 3. At 1 period horizon and for the D-Mark, the acceptance region would be:  $[-0.0014, 0.0230]$ , while the estimated  $dm$  is 0.0004. Hence, it is not possible to conclude that the random walk is a better description of the data than the economic model, even if it forecasts better. Table 2 reports the Bonferroni acceptance regions for all horizons (see the column labeled “*This paper’s quantiles*”) along with the value estimated in the data (see the column labeled “ $dm$ ”). The table shows that the same conclusion is valid at all horizons of prediction and for both currencies. For comparison, table 3 also reports the Diebold and Mariano (1995) and West (1996) quantiles.<sup>19</sup> Interestingly, note that if one used the Diebold and Mariano quantiles, one would often (wrongly) conclude that the random walk is a better description of the data because it forecasted significantly better. This would happen also if one used West’s quantiles for the D-Mark/\$ data.<sup>20</sup>

We now turn to the empirical evidence on the Monetary model. Table 4 reports the results from replicating Meese and Rogoff’s work for the monetary model. It is straightforward to notice that the same pattern observed for the RIRP arises here as well. Table 4 also reports the Bonferroni quantiles, along with the Diebold and Mariano (1995) and West (1996) quantiles. Again, we conclude that we cannot reject the null hypothesis.

---

are kernel smoothed by a normal kernel. Monte Carlo simulations of the  $dm$  statistic showed that the upper quantiles increase very quickly for near explosive processes, so we simulated only non-explosive processes.

<sup>19</sup>In practice,  $h$ -steps ahead forecasts will be at least  $(h-1)$  serially correlated. Hence, all covariances are estimated by using a Newey-West estimator, following the suggestions in West (1996). The bandwidth used in this paper is  $4P^{2/9}$ , which gives an effective symmetric bandwidth of approximately 9-10 lags. The results do not change substantially if one chooses a bigger/smaller bandwidth. If variables are stationary, quadratic loss functions imply that parameter estimation error is asymptotically irrelevant. However, in the framework of this paper it may play a role, because of the high persistence, so we report the result of the West test for comparison.

<sup>20</sup>Looking back to figure 4, last panel, it is clear that the forecasts of the model are farther away from the true future exchange rate than those of the random walk. We are not disputing the fact that a random walk forecast exchange rates well. What we show is that the same pattern could be observed *even if the economic model were the true DGP*, if the variables are highly persistent and the model (imprecisely) estimates the persistence.

Insert Tables 4 and 5

In order to check the robustness of the results to the estimation methods, we estimated the RIRPC by a VAR of order one. Qualitatively similar results hold (see table 5).

Bonferroni tests may have low power. Thus, the finding that we are unable to reject the hypothesis of equal predictive ability may be due to the low power of the test. To explore this issue, we perform a simple Monte Carlo experiment, calibrated on the actual data for given values of “**C**”. Figure 6 reports the results for the RIRPC and Monetary models for the D-Mark/\$ exchange rate ( $c_1 = c_2 = 5$  in the former and  $c_1 = c_2 = 10$  in the latter; these values belong to the confidence sets for **C**). Both figures plot the power functions of two nominal 5% tests: the test proposed in this paper and the (infeasible) test that uses the true values of the nuisance parameters. As the serial correlation increases (i.e.  $c$  moves towards zero or negative values), the power to reject the null hypothesis of equal predictive ability (in favor of the alternative that the random walk forecasts better) goes to one. As expected, the power of the feasible test is lower, which reflects the cost of the lack of knowledge on the roots, but the differences in power are not huge.

Insert Figure 6

## 7. Conclusions

The main result of this paper is the following: the fact that out-of-sample forecasts of a simple random walk outperform those of an economic model does not necessarily mean that the random walk is a better description of the data. The paper proposes a test for equal predictive ability that is robust to the presence of highly persistent variables, and generalizes the results in Diebold and Mariano (1995), West (1996) and Corradi, Swanson and Olivetti (1998). As an application, this paper reconsiders a well-known empirical fact uncovered by Meese and Rogoff (1983a and 1988), namely that a simple random walk model forecasts future nominal and real exchange rates better than theoretical economic models. The empirical analysis suggests that the results of Meese and Rogoff are not surprising, even assuming that the economic model is true. It is of course still true that the economic models do badly in terms of point forecasts (relative to the random walk).

The Bonferroni method used in this paper may deliver wide acceptance regions. Thus, the finding that it is not possible to reject the null hypothesis might be due to the low power of the test statistic. While this paper documented the existence of size distortions and a possible way to circumvent this problem, an important task for future research is to design tests with better power properties.

The result of this paper apply to situations in which the researcher is interested in predictions at long horizons in the presence of small sample sizes. This situation is quite common in finance and international macroeconomics (e.g. Mark (1995), Kilian (1999), Berkowitz and Giorgianni (2001)). Diebold and Kilian (2000) found evidence that unit root pre-tests are likely to improve the forecasting accuracy relative to forecasts from models in levels. Their Monte Carlo simulations suggest that pretesting may induce to select the incorrect forecasting model when the forecast horizon is large relative to the available sample size. Hence, their findings are similar to those in the present paper. Other recent works related to the present paper are Baillie and Bollerslev (2000) and Maynard and Phillips (2001), who suggest that another well-known empirical puzzle in international economics (the forward discount anomaly) might be explained by persistent variance or long memory. The findings of the above papers and the present one seem to suggest that this new strand of research, that focuses on a careful investigation of the properties of the data in finite samples, seems promising and exciting.

## References

- Baillie, Richard, and Bollerslev, Tim, “The forward premium anomaly is not as bad as you think”, *Journal of International Money and Finance* 19, 2000, pp. 471-488.
- Berkowitz, Jeremy, and Giorgianni, Lorenzo, “Long-Horizon Exchange Rate Predictability?”, *The Review of Economics and Statistics*, 83(1), February 2001, pp. 81-91.
- Cavanagh, Christopher L., Elliott, Graham, and Stock, James H., “Inference in Models with nearly Integrated Regressors”, *Econometric Theory* 11, 1995, pp. 1131-47.
- Clark, Todd, and McCracken, Michael, “Evaluating Long-Horizon Forecasts”, *mimeo*, December 2001(a).
- Clark, Todd, and McCracken, Michael, “Forecast-Based Model Selection in the Presence of Structural Break”, *mimeo*, December 2001(b).
- Clark, Todd, and McCracken, Michael, “Tests of equal forecast accuracy and encompassing for nested models”, *Journal of Econometrics* 105, 2001(c), pp. 85-110.
- Corradi, Valentina, Norman Swanson and Claudia Olivetti, “Predictive Ability with Cointegrated Variables”, *Journal of Econometrics* 104(2), September 2001, pp. 315-58.
- Diebold, Francis, and Mariano, Roberto, “Comparing Predictive Accuracy”, *Journal of Business Economics and Statistics*, July 1995.
- Diebold, Francis, and Kilian, Lutz, “Unit root tests are useful for selecting forecasting models”, *Journal of Business and Economic Statistics* 18(3), July 2000, pp. 265-73.
- Diebold, Francis, and Kilian, Lutz, “Measuring predictability: theory and macroeconomic applications”, *Journal of Applied Econometrics* 16(6), November-December 2001, pp. 657-69.
- Elliott, Graham, “On the robustness of cointegration methods when regressors almost have unit roots”, *Econometrica* 66(1), January 1998, pp. 149-58.
- Froot, Kenneth, and Rogoff, Kenneth, “Perspectives on PPP and Long-run Real Exchange Rates”, in Grossman, Gene M., and Rogoff, Kenneth, eds., *Handbook of International Economics*, vol. III, North Holland, 1995.
- Hamilton, James D., *Time Series Analysis*, Princeton University Press, 1994.
- Kemp, Gordon, “The behavior of forecast errors from a nearly integrated AR(1) model as both sample size and forecast horizon become large”, *Econometric Theory*, 15, 1999, pp. 238-256.
- Kilian, Lutz, “Exchange Rates and Monetary Fundamentals: What Do We Learn From Long-Horizon Regressions?”, *Journal of Applied Econometrics* 14(5), September-October 1999.
- Mark, Nelson C., “Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability”, *American Economic Review*, 85(1), March 1995, pp. 201-218.
- Maynard, Alex, and Phillips, Peter C. B., “Rethinking an Old Empirical Puzzle: Econometric

Evidence on the Forward Discount Anomaly”, *Journal of Applied Econometrics*, 16(6), November-December 2001, pp. 671-708.

McCracken, Michael W., “Asymptotics for Out of Sample Tests of Causality”, *mimeo*, Louisiana State University, 1999.

Meese, Richard, and Rogoff, Kenneth, “Exchange rate models of the seventies. Do they fit out of sample?”, *Journal of International Economics* 14, 1983(a).

Meese, Richard, and Rogoff, Kenneth, “The out-of-sample failure of empirical exchange rate models: Sampling error or mis-specification?”, in Frenkel, Jacob, ed., *Exchange rates and international macroeconomics*, University of Chicago Press, Chicago, 1983(b).

Meese, Richard, and Rogoff, Kenneth, “Was it real? The exchange rate-interest differential relation over the modern floating-rate period”, *The Journal of Finance* 43(3), September 1988.

Newey, Whitney, and West, Kenneth, “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix”, *Econometrica* 55, 1987, pp. 703-708.

Phillips, Peter C. B., “Towards a Unified Asymptotic Theory for Autoregression”, *Biometrika*, 74(3), 1987, pp.535-47.

Phillips, Peter C. B., “Regression Theory for Near-Integrated Time Series”, *Econometrica*, 56(5), September 1988 pp. 1021-1043.

Phillips, Peter C. B., “Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VARs”, *Journal of Econometrics*, 1998.

Richardson, Matthew and James Stock, “Drawing Inferences from Statistics Based on Multiyear Returns”, *Journal of Financial Economics* 25, 1989, pp. 323-348.

Rossi, Barbara, “Optimal tests for model selection with underlying parameter instability”, Duke Economics Working Paper 02-05, 2002.

Sims, Christopher A., Stock, James H., and Watson, Mark W., “Inference in linear time series models with some unit roots”, *Econometrica*, 58(1), January 1990, pp. 113-144.

Stock, James H., “VAR, Error Correction and Pretest Forecasts at Long Horizons”, *Oxford Bulletin of Economics and Statistics*, 58(4), 1996.

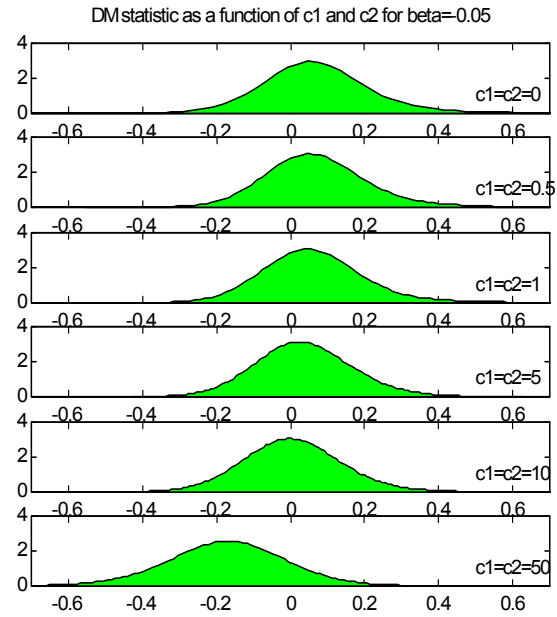
Stock, James H., “Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series”, *Journal of Monetary Economics*, 28 1991, pp. 435-459.

Stock, James H., and Watson, Mark W., “Testing for Common Trends”, *Journal of the American Statistical Association*, 83(404), December 1988, pp. 1097-1107.

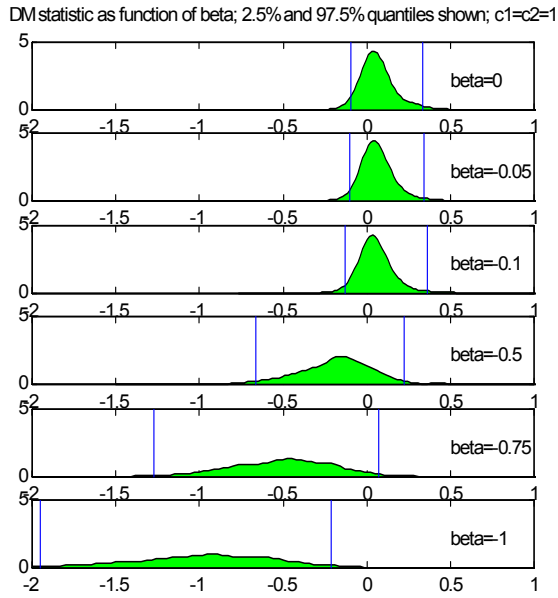
Stock, James H., and Watson, Mark W., “Confidence sets in regressions with highly serially correlated regressors”, *mimeo*, December 1996.

West, Kenneth, “Asymptotic Inference about Predictive Ability”, *Econometrica*, September 1996.

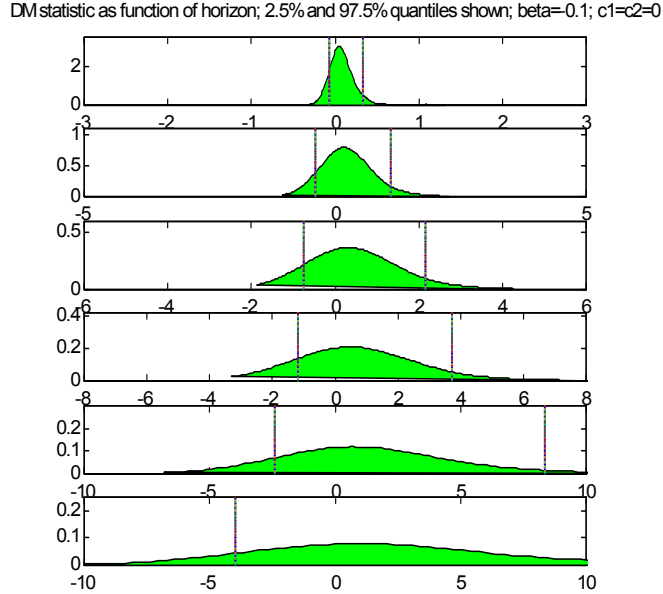
## Figures



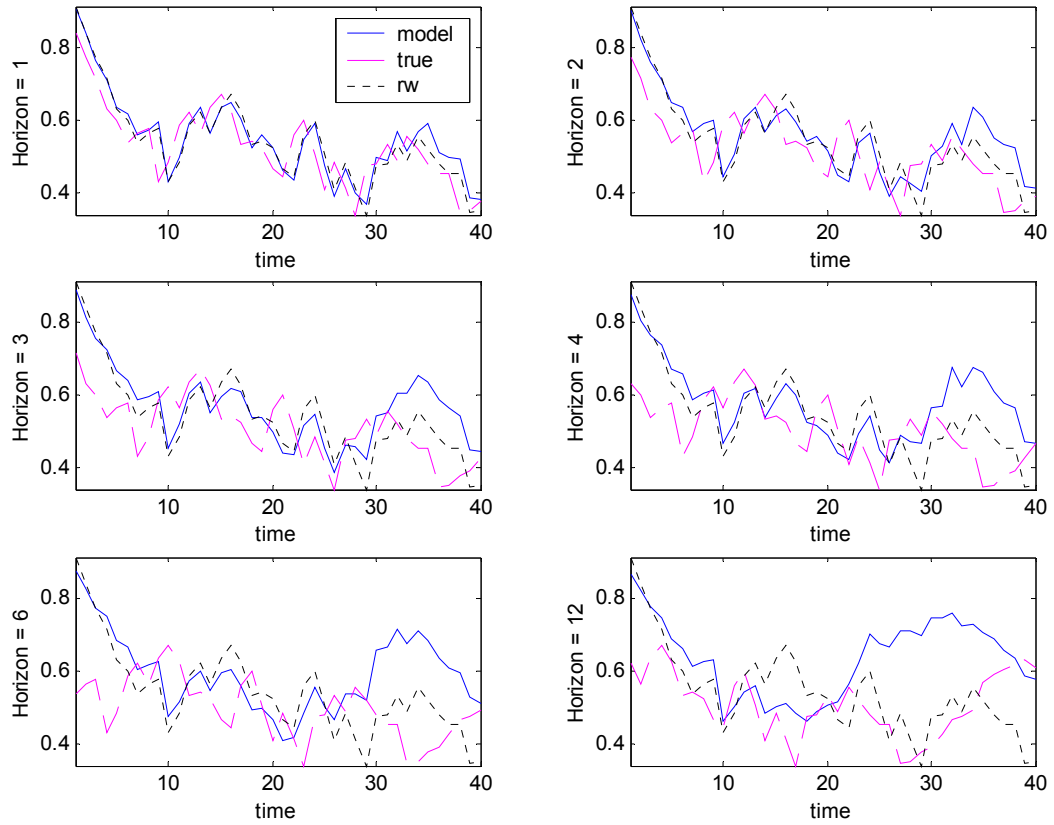
**Figure 1.** Simulated distribution of the estimated  $dm$  test statistic under the specified parameter values. Estimates are obtained in rolling samples. The horizon of prediction is 1 and the sample size 70.



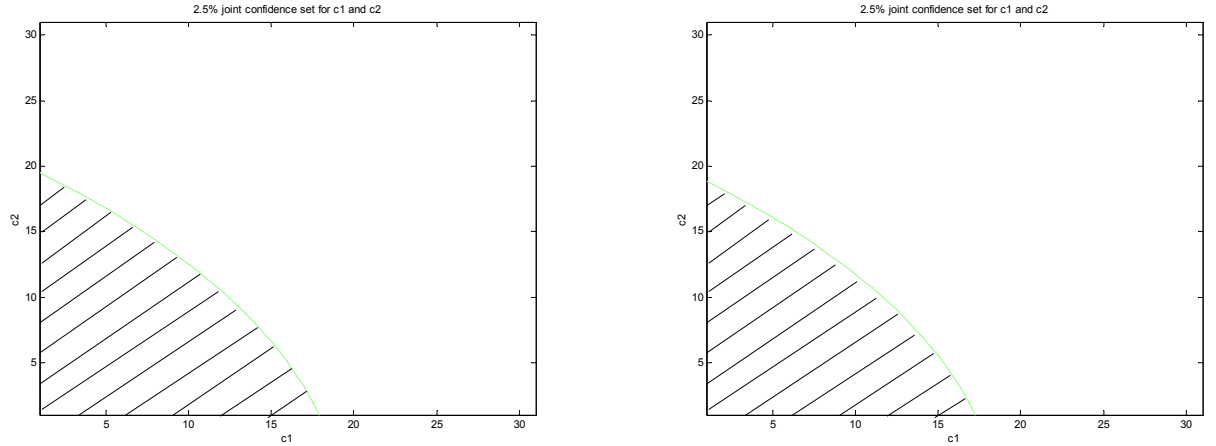
**Figure 2.** Figure 2 shows the simulated distribution of the estimated  $dm$  test statistic under the specified parameter values. Estimates are obtained in rolling samples. The horizon of prediction is 1 and the sample size 70. The vertical lines represent the 2.5% and 97.5% quantiles of the distribution.



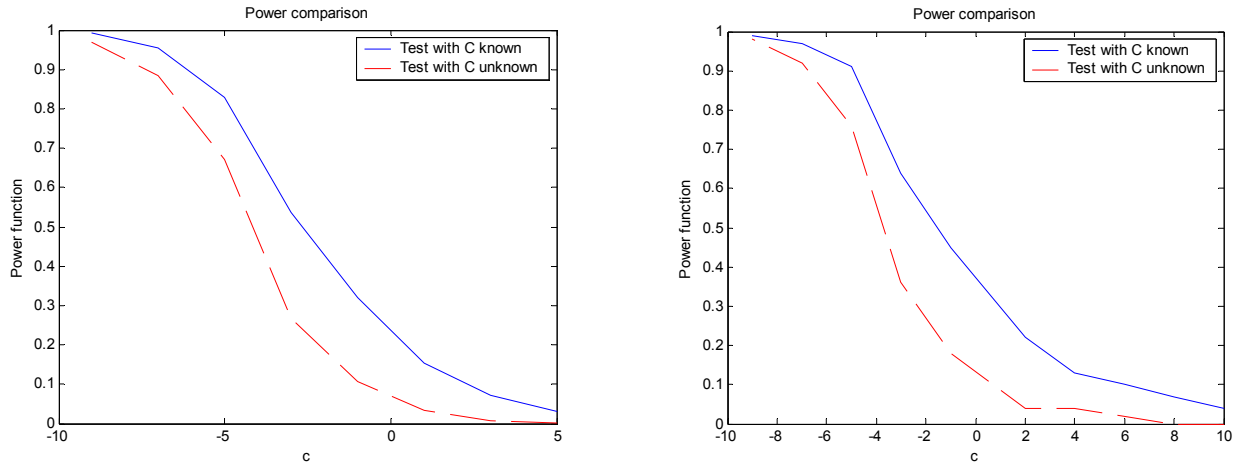
**Figure 3.** Simulated distribution of the estimated  $dm$  test statistic as function of the horizon of prediction under the specified parameter values. Estimates obtained in rolling samples. The horizon of prediction varies from 1 (top panel) to 2,3,4,6, 8 (bottom panel).



**Figure 4.** Comparison of predictions at various horizons for the D-mark/\$ RIRP model. The dashed line represents the true bilateral nominal exchange rate, the solid line represents the monetary model's predictions and the dotted line represents the random walk's predictions. Predictions of a random walk with drift model are (almost) indistinguishable from those of the random walk for every horizon and so are not reported.



**Figure 5.** The figure plots the 97.5% joint confidence set for  $(c_1, c_2)$ . The left panel reports the confidence set for the D-mark/\$ (the value of the test statistic in the data is  $-6.87$ ) and the right panel reports the confidence set for the Yen/\$ (the value of the test statistic in the data is  $-6.60$ ). The estimated model for both bilateral currencies is the RIRP model.



**Figure 6.** Power comparison of the feasible test used in this paper (dotted line) and the unfeasible test that knows the values of the nuisance parameters (solid line), as a function of the value of  $c$  ( $= c_1 = c_2$ ) for the rirpc (left panel) and monetary (right panel) models for D-Mark/\$ exchange rate. The simulations are calibrated on the actual, estimated parameter values other than  $c$ . The number of Monte Carlo replications is 2000.

## Tables

**Table 1.** (2.5%,5%,10%,50%,90%,95%,97.5%) percentiles of the simulated distribution of  $dm$  for figures 1 and 3.

	2.5%	5%	10%	50%	90%	95%	97.5%
<b>Figure 1</b>							
c=0	-0.1074	-0.0723	-0.0338	0.0554	0.2118	0.3079	0.3863
c=0.5	-0.0867	-0.0590	-0.0370	0.0524	0.1919	0.2684	0.3409
c=1	-0.1087	-0.0709	-0.0354	0.0497	0.1945	0.2682	0.3206
c=5	-0.1355	-0.0938	-0.0603	0.0213	0.1627	0.2077	0.2653
c=10	-0.1594	-0.1246	-0.0946	-0.0037	0.1087	0.1519	0.2085
c=50	-0.4866	-0.4165	-0.3535	-0.1698	-0.0131	0.0312	0.0801
c=-0.5	-0.1920	-0.1274	-0.0664	0.0536	0.2483	0.3395	0.4778
c=-1	-0.8738	-0.6842	-0.3459	0.0232	0.2860	0.4278	0.5697
<b>Figure 3</b>							
h=1	-0.1279	-0.0817	-0.0536	0.0431	0.1853	0.2484	0.3232
h=2	-0.3852	-0.2992	-0.1850	0.1217	0.5976	0.8002	1.0121
h=3	-0.8355	-0.6254	-0.4546	0.2066	1.2077	1.6550	2.1497
h=4	-1.3279	-1.1198	-0.7673	0.2941	1.9911	2.8034	3.4388
h=6	-2.8943	-2.2176	-1.6637	0.4309	3.9890	5.7437	8.4786
h=8	-4.9408	-3.6960	-2.7608	0.5672	6.5385	9.6468	16.0961

Note. In the table that refers to figures 1, “ $c = (.)$ ” means “ $c_1 = c_2 = (.)$ ”, as all simulations were performed for symmetric values of  $c_1$  and  $c_2$ . “ $h$ ” is the horizon of prediction.

**Table 2.** The Real Interest Rate Parity model

Currency	MSE $_h^m$	MSE $_h^{rw}$	This paper's	Diebold-Mariano	West	$dm$
$h$			quantiles	quantiles	quantiles	
D-Mark/\$						
1	.0049	.0045	(-.0014; .0230)	(-.0011; .0011)	(-.0182; .0182)	.0005
2	.0101	.0090	(-.0041; .0661)	(-.0031; .0031)	(-.0353; .0353)	.0011
3	.0136	.0112	(-.0078; .0989)	(-.0054; .0054)	(-.0511; .0511)	.0023
4	.0184	.0146	(-.0113; .1330)	(-.0082; .0082)	(-.0638; .0637)	.0038
6	.0282	.0209	(-.0289; .2052)	(-.0127; .0127)	(-.0871; .0871)	.0073
12	.0317	.0179	(-.0759; .3597)	(-.0211; .0211)	(-.1289; .1289)	.0138
Yen/\$						
1	.0051	.0047	(-.0012; .0147)	(-.0004; .0004)	(-.0244; .0243)	.0004
2	.0113	.0099	(-.0034; .0438)	(-.0012; .0012)	(-.0453; .0452)	.0014
3	.0149	.0121	(-.0060; .0788)	(-.002; .0020)	(-.0649; .0648)	.0028
4	.0200	.0157	(-.0090; .1104)	(-.0031; .0031)	(-.0833; .0833)	.0042
6	.0339	.0266	(-.0161; .1561)	(-.0061; .0061)	(-.1108; .1108)	.0072
12	.0599	.0514	(-.0440; .2188)	(-.0130; .0130)	(-.1710; .1710)	.0085

Note. The table reports statistics for the RIRP model for the two bilateral exchange rates. The horizon  $h$  is in quarters. The statistics showed in this table are: the MSE of the monetary model (MSE $_h^m$ ), the MSE of the random walk (MSE $_h^{rw}$ ), the difference MSE $_h^m$  and MSE $_h^{rw}$  ( $dm$ ). The fourth column contains the quantiles for testing equal predictive ability by using the methods described in this paper. The fifth and sixth columns report, respectively, the (2.5;97.5)% quantiles for testing the same null hypothesis by using the Diebold and Mariano (1995) and West (1996) test statistics.

**Table 3.** Lower (.0125) and upper (.9875) critical values of the  $dm$  statistic

$c_1$	$c_2$	$dm_L^{h=1}$	$dm_U^{h=1}$	$dm_L^{h=3}$	$dm_U^{h=3}$	$dm_L^{h=6}$	$dm_U^{h=6}$	$dm_L^{h=12}$	$dm_U^{h=12}$
0	0	-0.0012	0.023	-0.0056	0.0986	-0.0266	0.2047	-0.0759	0.3494
0	4	-0.0011	0.0229	-0.0053	0.0988	-0.0225	0.205	-0.0648	0.3538
0	16	-0.001	0.023	-0.0045	0.0989	-0.0149	0.2051	-0.044	0.3597
4	0	-0.0012	0.0165	-0.0062	0.0698	-0.0272	0.1406	-0.0735	0.2286
4	4	-0.0011	0.0164	-0.0059	0.07	-0.0232	0.1409	-0.0625	0.233
16	0	-0.0014	0.0048	-0.0081	0.0159	-0.0292	0.023	-0.0667	0.0177

Note:  $c_1$  and  $c_2$  are a subset of the simulated values of the local-to-unity parameters.  $dm_L^{h=j}$  and  $dm_U^{h=j}$  are the corresponding values of the Lower (.0125) and Upper (.9875) quantiles of the  $dm$  statistic at horizon  $j$ , for  $j=1,3,6,12$ . The model is the RIRP condition and the data are for the D-Mark/\$ exchange rate.

**Table 4.** The Monetary model.

Currency	$MSE_h^m$	$MSE_h^{rw}$	This paper's	Diebold-Mariano	West	$dm$
$h$			quantiles	quantiles	quantiles	
D-Mark/\$						
1	.0089	.0051	(-.0012; .0357)	(-.0033; .0033)	(-.0031; .0031)	.0037
2	.0215	.0106	(-.0044; .1050)	(-.0096; .0096)	(-.0090; .0090)	.011
3	.0364	.0136	(-.0089; .1772)	(-.0178; .0178)	(-.0175; .0175)	.0228
4	.0571	.0180	(-.0140; .2574)	(-.0266; .0266)	(-.0265; .0265)	.0391
6	.1023	.0278	(-.0257; .3782)	(-.0474; .0474)	(-.0476; .0476)	.0746
12	.2068	.0272	(-.0793; .6236)	(-.1885; .1885)	(-.1868; .1868)	.1797
Yen/\$						
1	.0063	.0047	(-.0008; .0228)	(-.0022; .0022)	(-.0136; .0136)	.0016
2	.0147	.0106	(-.0031; .0655)	(-.0065; .0065)	(-.0249; .0249)	.0041
3	.0214	.0146	(-.0061; .1167)	(-.0104; .0104)	(-.0367; .0367)	.0068
4	.0298	.0203	(-.0092; .1608)	(-.0147; .0147)	(-.0458; .0458)	.0095
6	.0470	.0327	(-.0166; .2366)	(-.0230; .0230)	(-.0550; .0550)	.0143
12	.0894	.0618	(-.0558; .3854)	(-.0357; .0357)	(-.0728; .0728)	.0276

**Table 5.** The VAR model.

Currency	$MSE_h^m$	$MSE_h^{rw}$	This paper's	Diebold-Mariano	West	$dm$
$h$			quantiles	quantiles	quantiles	
D-Mark/\$						
1	.0047	.0045	(-.0018; .9064)	(-.0004; .0004)	(-.0176; .0176)	.0002
2	.0097	.0090	(-.0071; 2.329)	(-.0013; .0013)	(-.0191; .0191)	.0007
3	.0124	.0112	(-.0145; 3.736)	(-.0024; .0024)	(-.0178; .0178)	.0012
4	.0165	.0146	(-.0190; 4.811)	(-.0037; .0037)	(-.0129; .0129)	.0019
6	.0239	.0209	(-.0254; 5.915)	(-.0060; .0060)	(-.0115; .0115)	.0030
12	.0190	.0179	(-.0577; 6.772)	(-.0105; .0105)	(-.0396; .0396)	.0011
Yen/\$						
1	.0047	.0047	(-.0007; .0010)	(-.0002; .0002)	(-.0302; .0302)	.0000
2	.0098	.0099	(-.0022; .0038)	(-.0009; .0009)	(-.0213; .0213)	-.0001
3	.0114	.0121	(-.0041; .0077)	(-.0018; .0018)	(-.0139; .0139)	-.0007
4	.0149	.0157	(-.0064; .0114)	(-.0034; .0034)	(-.0092; .0092)	-.0008
6	.0269	.0266	(-.0107; .0221)	(-.0085; .0085)	(-.0120; .0120)	.0003
12	.0539	.0514	(-.0245; .0597)	(-.0248; .0248)	(-.0667; .0667)	.0025

Notes to tables 4 and 5: as per Table 2.

## Appendix

### Proof of the derivation of (15).

For simplicity, the notation will abstract from the dependence on  $\Psi$  and on the information set, which is  $\{y_s\}_{s=1}^t$  here. By using the Wiener-Kolmogorov prediction formula (see Hamilton (1994), chp. 4), the DGP can be rewritten as:

$$w_{t+h} = e_{t+h} + \widehat{w}_{t+h}$$

where:

$$\begin{aligned} e_{t+h} &\equiv \frac{I + \Phi L + \dots + \Phi^{h-1} L^{h-1}}{L^h} \epsilon_t + \frac{(I + \Phi L + \dots + \Phi^{h-2} L^{h-2}) \Theta_1 L}{L^h} \epsilon_t + \\ &\quad + \dots + \frac{(I + \Phi L + \dots + \Phi^{h-q} L^{h-q}) \Theta_q L^q}{L^h} \epsilon_t \\ \widehat{w}_{t+h} &\equiv \left( \frac{\Phi^h L^h + \Phi^{h+1} L^{h+1} + \dots}{L^h} + \frac{\Phi^{h-1} L^{h-1} + \Phi^h L^h + \dots}{L^h} \Theta_1 L + \dots + \frac{\Phi^{h-q} L^{h-q} + \dots}{L^h} \Theta_q L^q \right) \epsilon_t \end{aligned}$$

Rewrite the unforecastable component  $e_{t+h}$  as:

$$\begin{aligned} e_{t+h} &= \epsilon_{t+h} + \Phi \epsilon_{t+h-1} + \Phi^2 \epsilon_{t+h-2} + \dots + \Phi^{h-1} \epsilon_{t+1} + \\ &\quad + \Theta_1 \epsilon_{t+h-1} + \Phi \Theta_1 \epsilon_{t+h-2} + \dots + \Phi^{h-2} \Theta_1 \epsilon_{t+1} + \\ &\quad + \Theta_2 \epsilon_{t+h-2} + \dots + \Phi^{h-3} \Theta_2 \epsilon_{t+1} + \dots \\ &= \epsilon_{t+h} + \Phi (I + \Phi^{-1} \Theta_1) \epsilon_{t+h-1} + \Phi^2 (I + \Phi^{-1} \Theta_1 + \Phi^{-2} \Theta_2) \epsilon_{t+h-2} + \\ &\quad \dots + \Phi^{q-1} (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q+1} \Theta_{q-1}) \epsilon_{t+h-q+1} + \\ &\quad + \Phi^q (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h-q} + \Phi^{q+1} (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h-q-1} + \\ &\quad + \dots + \Phi^{h-1} (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+1} \\ &= \sum_{j=0}^{h-1} \Phi^j (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h-j} \\ &\quad - (\Phi^{-1} \Theta_1 + \Phi^{-2} \Theta_2 + \Phi^{-3} \Theta_3 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h} - \\ &\quad - \Phi ( \quad \quad \quad \Phi^{-2} \Theta_2 + \Phi^{-3} \Theta_3 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h-1} - \\ &\quad - \dots \quad \quad \quad - \Phi^{q-1} (\Phi^{-q} \Theta_q) \epsilon_{t+h-q+1} = \\ &= \sum_{j=0}^{h-1} \Phi^j (I + \Phi^{-1} \Theta_1 + \dots + \Phi^{-q} \Theta_q) \epsilon_{t+h-j} - \sum_{j=0}^{q-1} \Phi^j \sum_{s=j+1}^q \Phi^{-s} \Theta_s \epsilon_{t+h-j} \end{aligned}$$

Then consider the forecastable component:

$$\begin{aligned} \widehat{w}_{t+h} &= \Phi^h \epsilon_t + \Phi^{h+1} \epsilon_{t-1} + \Phi^{h+2} \epsilon_{t-2} + \dots + \Phi^{h+q-1} \epsilon_{t-q+1} + \dots \\ &\quad + \Phi^{h-1} \Theta_1 \epsilon_t + \Phi^h \Theta_1 \epsilon_{t-1} + \Phi^{h+1} \Theta_1 \epsilon_{t-2} + \dots + \Phi^{h+q-2} \Theta_1 \epsilon_{t-q+1} + \dots \\ &\quad \dots \dots \dots \\ &\quad + \Phi^{h-q} \Theta_q \epsilon_t + \Phi^{h-q+1} \Theta_q \epsilon_{t-1} + \Phi^{h-q+2} \Theta_q \epsilon_{t-2} + \dots + \Phi^{h-1} \Theta_q \epsilon_{t-q+1} + \dots \end{aligned}$$

and compare it with:

$$\begin{aligned} \Phi^h w_t &= \Phi^h \sum_{j=0}^{\infty} \Phi^j u_{t-j} = \\ &= \Phi^h \epsilon_t + \Phi^h \Theta_1 \epsilon_{t-1} + \Phi^h \Theta_2 \epsilon_{t-2} + \dots + \Phi^h \Theta_{q-1} \epsilon_{t-q+1} + \dots \end{aligned}$$

$$\begin{aligned}
& +\Phi^{h+1}\epsilon_{t-1} + \Phi^{h+1}\Theta_1\epsilon_{t-2} + \dots + \Phi^{h+1}\Theta_{q-2}\epsilon_{t-q+1} + \dots \\
& \quad +\Phi^{h+2}\epsilon_{t-2} + \dots + \Phi^{h+q-1}\epsilon_{t-q+1} + \dots = \\
= & \widehat{w}_{t+h} - \Phi^h (\Phi^{-1}\Theta_1 + \Phi^{-2}\Theta_2 + \dots + \Phi^{-q}\Theta_q) \epsilon_t - \\
& -\Phi^{h+1} (\Phi^{-2}\Theta_2 + \Phi^{-3}\Theta_3 + \dots + \Phi^{-q}\Theta_q) \epsilon_{t-1} - \\
& -\dots - \Phi^{h+(q-1)} (\Phi^{-q}\Theta_q) \epsilon_{t-q+1} = \\
= & \widehat{w}_{t+h} - \Phi^h \sum_{j=0}^{q-1} \Phi^j \sum_{s=j+1}^q \Phi^{-s} \Theta_s \epsilon_{t-j} \\
\text{Thus } w_{t+h} = & \sum_{j=0}^{h-1} \Phi^j (I + \Phi^{-1}\Theta_1 + \dots + \Phi^{-q}\Theta_q) \epsilon_{t+h-j} + \Phi^h w_t + a_{t,h} \text{ where:}
\end{aligned}$$

$$a_{t,h} \equiv \Phi^h \sum_{j=0}^{q-1} \Phi^j \sum_{s=j+1}^q \Phi^{-s} \Theta_s \epsilon_{t-j} - \sum_{j=0}^{q-1} \Phi^j \sum_{s=j+1}^q \Phi^{-s} \Theta_s \epsilon_{t+h-j} = (\Phi^h - L^{-h}) \sum_{j=0}^{q-1} \Phi^j \sum_{s=j+1}^q \Phi^{-s} \Theta_s \epsilon_{t-j} \quad (32)$$

When the variables are highly persistent,  $(I + \Phi^{-1}\Theta_1 + \dots + \Phi^{-q}\Theta_q)$  can be approximated by  $\Theta(I)$  and  $a_{t,h}$  is of order smaller than the other components and, thus, it is asymptotically irrelevant (i.e. it is  $O_p(1)$  whereas all the other components are  $O_p(\sqrt{T})$ ). Hence:

$$w_{t+h} = \sum_{i=0}^{h-1} \Phi^i \Theta(I) \epsilon_{t+h-i} + \Phi^h w_t + o_p(T^{1/2}) \quad (33)$$

### Proof that VAR estimation satisfies Assumption (B)

Assume that the DGP (5) can alternatively be represented as a finite order  $VAR(p)$  process in levels (i.e.  $\Theta(L)^{-1} \simeq (I - \bar{\Theta}_1 L - \bar{\Theta}_2 L^2 - \dots - \bar{\Theta}_{p-1} L^{p-1} \equiv \bar{\Theta}(L))$ ):

$$(I - \bar{\Theta}_1 L - \bar{\Theta}_2 L^2 - \dots - \bar{\Theta}_{p-1} L^{p-1}) (I - \Phi L) (By_t - \bar{y} - Dt) = \epsilon_t \quad (34)$$

The reduced form VAR satisfies:

$$(I - \Theta_1^* L - \Theta_2^* L^2 - \dots - \Theta_{p-1}^* L^{p-1}) (I - \Phi^* L) (y_t - \bar{y}^* - D^* t) = \epsilon_t^* \quad (35)$$

where  $\bar{y}^* \equiv B^{-1}\bar{y}$ ,  $D^* \equiv B^{-1}D$ ,  $\tilde{d}^* \equiv B^{-1}\tilde{d}$ ,  $\epsilon_t^* \equiv B^{-1}\epsilon_t$  and all other “starred” parameters satisfy  $\Theta_j^* \equiv B^{-1}\bar{\Theta}_j B \forall j = 1, \dots, (p-1)$  and  $\Phi^* \equiv B^{-1}\Phi B$ . Also, let  $w_t^* \equiv B^{-1}w_t = y_t - \bar{y}^* - D^* t$  and  $\tilde{\Delta} \equiv (I - \Phi^* L)$ . Thus, the DGP written in terms of the canonical regressors is:

$$y_t = \bar{y}^* + D^* t + \Phi^* w_{t-1}^* + \sum_{j=1}^{p-1} \Theta_j^* \tilde{\Delta} w_{t-j}^* \quad (36)$$

We assume that the researcher estimates a finite order  $VAR(p)$ :

$$y_t = \mu_0 + \mu_1 t + \sum_{j=1}^p J_j y_{t-j} + \epsilon_t \quad (37)$$

which we can rewrite in its companion form:

$$Y_t = I_1 (\mu_0 + \mu_1 t) + \underbrace{\begin{pmatrix} J_1 & J_2 & \dots & J_p \\ I & 0 & \dots & 0 \\ & & \dots & \\ 0 & \dots & I & 0 \end{pmatrix}}_{\equiv \bar{A}} Y_{t-1} + I_1 \epsilon_t \quad (38)$$

where  $Y_{t-1}, Y_t \equiv (y'_t, y'_{t-1}, \dots, y'_{t-p+1})'$ ,  $I_1 \equiv \begin{pmatrix} I & 0 & 0 & \dots & 0 \end{pmatrix}'$  and  $\mu_0, \mu_1$  and  $J_i$  for  $i = 1..p$  are, respectively,  $(n \times 1)$ ,  $(n \times 1)$  and  $(n \times n)$  matrices of coefficients. If not specified otherwise (by a subscript that describes the dimension),  $I$  is an  $(n \times n)$  identity matrix and  $0$  is an  $(n \times n)$  matrix of zeros.

To evaluate parameter estimation error at long horizons, which involves  $\bar{A}^h$ , we use Phillips (1998) device to rewrite the system as:

$$\tilde{Y}_t = I_2 (\mu_0 + \mu_1 t) + A \tilde{Y}_{t-1} + I_2 \epsilon_t \quad (39)$$

where  $\tilde{Y}_t \equiv (y'_t, \Delta y'_t, \dots, \Delta y'_{t-p+2})'$ ,  $I_2 \equiv \begin{pmatrix} I & I & 0 & \dots & 0 \end{pmatrix}'$  and:

$$A = \begin{pmatrix} J(1) & -\Sigma_{j=2}^p J_j & -\Sigma_{j=3}^p J_j & \dots & -J_p \\ J(1) - I & -\Sigma_{j=2}^p J_j & -\Sigma_{j=3}^p J_j & \dots & -J_p \\ 0 & I & 0 & \dots & \\ & & \dots & & \\ \dots & 0 & I & 0 & 0 \\ 0 & \dots & 0 & I & 0 \end{pmatrix} \quad (40)$$

In what follows, we will use  $A_{ij}$  to denote the  $(n \times n)$  sub-matrices of  $A$  in (40) (there are  $p^2$  of them) and  $A_{j:k,l:m}$  to denote the submatrix of  $A$  formed by taking submatrices with row subscripts  $j, j+1, \dots, k$  and column subscripts  $l, l+1, \dots, m$ . Note that the non-stationary components have been separated in the first left sub-matrix only,  $J(1)$ . Since  $J(1) = I + C_A T^{-1}$ ,  $C_A \equiv \Theta(I)C$ , then  $J(1) - I$  is negligible asymptotically and we can rewrite (40) as:

$$A \simeq \begin{pmatrix} J(1) & A_{1,2:p} \\ 0_{(p-1)n \times n} & A_{2:p,2:p} \end{pmatrix}$$

We use the device in Sims, Stock and Watson (1990) and Elliott (1998) to rewrite (34) so that the components have different orders of integration.

*Lemma 2. The canonical regression transformation*

*The mapping from the canonical regressors in (36) and the regressors used for estimation in (37)*

*is given by (see Stock (1991)):*

$$\begin{pmatrix} \tilde{\Delta}w_{t-1}^* \\ \tilde{\Delta}w_{t-2}^* \\ \dots \\ \tilde{\Delta}w_{t-p}^* \\ 1 \\ w_{t-1}^* \\ t \end{pmatrix} = \begin{pmatrix} \Phi^* & 0 & 0 & 0 & H_{1,p+1} & (I - \Phi^*) & -(I - \Phi^*)d^* \\ -(I - \Phi^*) & \Phi^* & 0 & 0 & H_{2,p+1} & (I - \Phi^*) & -(I - \Phi^*)d^* \\ \dots & \dots & \dots & 0 & \dots & \dots & \dots \\ -(I - \Phi^*) & -(I - \Phi^*) & \dots & \Phi^* & H_{p,p+1} & (I - \Phi^*) & -(I - \Phi^*)d^* \\ 0 & 0 & \dots & 0 & 1 & 0_{1 \times n} & 0_{1 \times 1} \\ 0 & 0 & \dots & 0 & -\bar{y}^* + d^* & I & -d^* \\ 0 & 0 & \dots & 0 & 0_{1 \times 1} & 0_{1 \times n} & 1 \end{pmatrix} \begin{pmatrix} \Delta y_{t-1} \\ \Delta y_{t-2} \\ \dots \\ \Delta y_{t-p} \\ 1 \\ y_{t-1} \\ t \end{pmatrix}$$

where  $H_{j,p+1} \equiv -(I - \Phi^*)(\bar{y}^* - jd^*) - \Phi^*d^*$ . Thus, the mapping on the parameters is given by:

$$\begin{pmatrix} A'_{12} \\ A'_{13} \\ \dots \\ A'_{1p+1} \\ \mu'_0 \\ J(1)' \\ \mu'_1 \end{pmatrix} = \begin{pmatrix} \Phi^{*'} & -(I - \Phi^*)' & \dots & -(I - \Phi^*)' & 0 & 0 & 0 \\ 0 & \Phi^{*'} & \dots & -(I - \Phi^*)' & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Phi^{*'} & \dots & (I - \Phi^*)' & -d^{*'}(I - \Phi^*)' \\ H'_{1,p+1} & H'_{2,p+1} & \dots & H'_{p,p+1} & 1 & (-\bar{y}^* + d^*)' & 0_{1 \times 1} \\ (I - \Phi^*)' & (I - \Phi^*) & \dots & (I - \Phi^*)' & 0_{n \times 1} & I & 0_{n \times 1} \\ -d^{*'}(I - \Phi^*)' & -d^{*'}(I - \Phi^*)' & \dots & -d^{*'}(I - \Phi^*)' & 0_{1 \times 1} & -d^{*'} & 1 \end{pmatrix} \begin{pmatrix} \Theta_1^{*'} \\ \Theta_2^{*'} \\ \dots \\ \Theta_p^{*'} \\ \bar{y}^{*'} \\ \Phi^{*'} \\ d^{*'} \end{pmatrix}$$

It follows that:<sup>21</sup>

$$\begin{aligned} \sqrt{T}(\hat{A}_{1k+1} - \Theta_k^*) &= O_p(1) \quad \forall k = 1, 2, \dots, p-1 \\ \sqrt{T}(\hat{\mu}_0 - 0) &= O_p(1) \\ T^{3/2}(\hat{\mu}_1 - 0) &= O_p(1) \\ T(\hat{J}(1) - \Phi^*) &= O_p(1) \end{aligned} \tag{41}$$

Note that  $\hat{J}(1)$  is the estimate of the unit root matrix  $\Phi^*$ , i.e.  $\hat{J}(1) = \hat{\Phi}^*$ , and  $\hat{A}_{1k+1}$  are the estimates of  $\Theta_k^*$ , i.e.  $\hat{A}_{1k+1} = \hat{\Theta}_k^*$ . It follows that the estimated coefficients (b)-(e) satisfy Assumption B when rewritten with “\*”. Note that the rates of convergence are the same as those in Sims, Stock and Watson (1990), for the case of exact  $I(1)$  variables.

<sup>21</sup>Let  $C^* \equiv B^{-1}CB$ . Recall that  $d^* = T^{-1/2}\tilde{d}^*$  and  $(I - \Phi^*) = C^*T^{-1}$  are of smaller order. Thus:  $\hat{\mu}_1 - C^*\Theta(1)^*d^* = \hat{\mu}_1 - C^*\Theta(1)^*T^{-1/2}\tilde{d}^* = \hat{\mu}_1 - \mu_1 + o_p(1)$  for  $\mu_1 = 0$  and  $\hat{\mu}_0 - C^*T^{-1}\Theta(1)^*\bar{y}^* + C^*T^{-1}\tilde{\Theta}^*(1)d^* + \Phi^*\Theta(1)^*d^* = \hat{\mu}_0 - C^*T^{-1}\Theta(1)^*\bar{y}^* + C^*T^{-3/2}\tilde{\Theta}^*(1)d^* + \Phi^*\Theta(1)^*T^{-1/2}\tilde{d}^* = \hat{\mu}_0 - \mu_0 + o_p(1)$ , where  $\mu_0 = 0$  and  $\tilde{\Theta}^*(1) = \sum_{j=1}^{p-1} \sum_{k=j+1}^p \Theta_k^*$ .

**Proof of (21)**

At long horizons, by iterating on (39), we find that the dynamics of  $y_{t+h}$  is described by:

$$y_{t+h} = I_1' \widetilde{Y}_{t+h} = I_1' \sum_{j=0}^{h-1} A^j I_2 (\mu_0 + \mu_1 (t+h-j)) + I_1' A^h \widetilde{Y}_t + e_{t+h}^m (\Psi | \mathfrak{S}_t) \quad (42)$$

where  $e_{t+h}^m (\Psi | \mathfrak{S}) \equiv I_1' \sum_{j=0}^{h-1} A^j I_2 \epsilon_{t+h-j}$ . Calculations show that:

$$A^h \simeq \begin{pmatrix} J(1)^h & \sum_{j=0}^{h-1} J(1)^{h-j-1} A_{1,2:p} A_{2:p,2:p}^j \\ 0_{(p-1)n \times n} & A_{2:p,2:p}^h \end{pmatrix}$$

As  $h \rightarrow \infty$ ,  $A_{2:p,2:p}^h \rightarrow 0_{(p-1) \times (p-1)}$ , as all eigenvalues of  $A_{2:p,2:p}$  are less than one in modulus, and  $I_1' \sum_{j=0}^{h-1} A^j I_2 \simeq \sum_{j=0}^{h-1} J(1)^j (I - \sum_{j=2}^p A_{1,j})^{-1}$ . In fact:

$$\begin{aligned} & I_1' \sum_{j=0}^{h-1} A^j I_2 \\ = & I_1' \begin{pmatrix} \sum_{j=0}^{h-1} J(1)^j & \sum_{j=0}^{h-1} J(1)^{j-1} \sum_{i=0}^{j-1} J(1)^{-i} A_{1,2:p} A_{2:p,2:p}^i \\ 0_{(p-1)n \times n} & \sum_{j=0}^{h-1} A_{2:p,2:p}^j \end{pmatrix} I_2 \\ = & \sum_{j=0}^{h-1} J(1)^j + \sum_{j=0}^{h-1} J(1)^{j-1} \sum_{i=0}^{j-1} J(1)^{-i} A_{1,2:p} A_{2:p,2:p}^i I_1 \\ \simeq & \sum_{j=0}^{h-1} J(1)^j + \sum_{j=0}^{h-1} J(1)^{j-1} A_{1,2:p} (I - A_{2:p,2:p}^i) (I - A_{2:p,2:p})^{-1} I_1 \\ = & \sum_{j=0}^{h-1} J(1)^j + \sum_{j=0}^{h-1} J(1)^{j-1} A_{1,2:p} (I - A_{2:p,2:p})^{-1} I_1 \\ & - \sum_{j=0}^{h-1} J(1)^{j-1} A_{1,2:p} A_{2:p,2:p}^i (I - A_{2:p,2:p})^{-1} I_1 \\ \simeq & \sum_{j=0}^{h-1} J(1)^j \left( I + A_{1,2:p} (I - A_{2:p,2:p})^{-1} \right) I_1 \\ = & \sum_{j=0}^{h-1} J(1)^j (I - \sum_{j=2}^p A_{1,j})^{-1} \end{aligned}$$

where the fifth passage follows from the fact that the last addend is of smaller order than the other components, and the last equality follows from direct calculations and the special structure of  $A$ .

Thus, the forecast error of the model,  $e_{t+h}^m (\widehat{\Psi} | \mathfrak{S}_t)$ , is the first row of:

$$\begin{aligned} & y_{t+h} - \widehat{J}(1)^h y_t - \widehat{J}(1)^{h-1} \widehat{A}_{1,2:p} \left( I - \widehat{A}_{2:p,2:p} \right)^{-1} (\Delta y'_t, \dots, \Delta y'_{t-p+2})' \\ & - \sum_{j=0}^{h-1} \widehat{J}(1)^j \left( I - \sum_{j=2}^p \widehat{A}_{1,j} \right)^{-1} (\widehat{\mu}_0 + (t+h-j) \widehat{\mu}_1) \end{aligned} \quad (43)$$

Notice however that the third element is of smaller order than the others, and from lemma 2,  $I - \sum_{j=2}^p \widehat{A}_{1,j}$  is a  $\sqrt{T}$ -consistent estimate of  $\Theta^*(1)$ . Thus:

$$e_{t+h}^m (\widehat{\Psi} | \mathfrak{S}_t) = y_{1,t+h} - i_1' \widehat{J}(1)^h y_t - i_1' \sum_{j=0}^{h-1} \widehat{J}(1)^j \Theta^*(1)^{-1} (\widehat{\mu}_0 + (t+h-j) \widehat{\mu}_1) + o_p(\sqrt{T}) \quad (44)$$

On the other hand, the population forecast error,  $e_{t+h}^m (\Psi | \mathfrak{S}_t)$ , is  $i_1' (w_{t+h}^* - \Phi^{*h} w_t^*)$ , which is the first row of:

$$y_{t+h} - \Phi^{*h} y_t - (I(t+h) - \Phi^{*h} t) d^* + o_p(\sqrt{T}) \quad (45)$$

Finally, the difference between (44) and (45) is:

$$\begin{aligned}
e_{t+h}^m(\Psi|\mathfrak{S}_t) - e_{t+h}^m(\widehat{\Psi}|\mathfrak{S}_t) &= i'_1 \left( \widehat{J}(1)^h - \Phi^{*h} \right) y_t + i'_1 \sum_{j=0}^{h-1} \widehat{J}(1)^j \Theta^* (1)^{-1} \widehat{\mu}_0 + \\
&\quad + i'_1 \sum_{j=0}^{h-1} \widehat{J}(1)^j (t+h-j) \Theta^* (1)^{-1} \widehat{\mu}_1 + \\
&\quad - i'_1 \left( I(t+h) - \Phi^{*ht} \right) d^* + o_p \left( \sqrt{T} \right) \\
&= i'_1 B^{-1} \left( B \widehat{J}(1)^h B^{-1} - \Phi^h \right) w_t + i'_1 B^{-1} \left( B \widehat{J}(1)^h B^{-1} - \Phi^h \right) Dt \\
&\quad + i'_1 B^{-1} \sum_{j=0}^{h-1} B \widehat{J}(1)^j B^{-1} \Theta (1)^{-1} B [\widehat{\mu}_0 + (t+h-j) \widehat{\mu}_1] - \\
&\quad - i'_1 B^{-1} \left( I(t+h) - B \widehat{J}(1)^h B^{-1} t \right) D + o_p \left( \sqrt{T} \right)
\end{aligned}$$

Thus:

$$T^{-1/2} \left( e_{t+h}^m(\Psi|\mathfrak{S}_t) - e_{t+h}^m(\widehat{\Psi}|\mathfrak{S}_t) \right) = i'_1 B^{-1} \left( B \widehat{J}(1)^h B^{-1} - \Phi^h \right) \left( T^{-1/2} w_t \right) + \widehat{\zeta}_{pe}^{\det} \quad (46)$$

where:

$$\begin{aligned}
\widehat{\zeta}_{pe}^{\det} &\equiv i'_1 B^{-1} T^{-1} \left( I(t+h) - B \widehat{J}(1)^h B^{-1} t \right) \widetilde{d} - i'_1 B^{-1} \left( B \widehat{J}(1)^h B^{-1} - \Phi^h \right) Dt + \\
&\quad - i'_1 B^{-1} \left( T^{-1} \sum_{j=0}^{h-1} B \widehat{J}(1)^j B^{-1} \right) \Theta (1)^{-1} B [T^{1/2} \widehat{\mu}_0 + T^{-1} (t+h-j) T^{3/2} \widehat{\mu}_1] + o_p(1)
\end{aligned} \quad (47)$$

Define  $V_{C_A}$  and  $\Lambda_{C_A}$  to be the eigenvector and eigenvalue matrices of  $C_A$ . Thus,  $B \widehat{J}(1)^h B^{-1} \rightarrow B V_{C_A} \mathbf{e}^{-\delta \Lambda_{C_A}} V_{C_A}^{-1} B^{-1} \equiv \mathbf{e}^{-\delta \widetilde{C}}$ . Also,  $\Phi^h \rightarrow \mathbf{e}^{-\delta C}$ . As  $h \rightarrow \infty$ , the first component in (46) is such that  $i'_1 B^{-1} \left( B \widehat{J}(1)^h B^{-1} - \Phi^h \right) \left( T^{-1/2} w_t \right) \Rightarrow i'_1 B^{-1} \left( \mathbf{e}^{-\delta \widetilde{C}} - \mathbf{e}^{-\delta C} \right) \Lambda_{J_c}(\tau)$  by lemma 1.  $\widehat{\zeta}_{pe}^{\det}$ , which reflects the contribution of parameter estimation error in the deterministic trend components, is  $O_p(1)$ ; denote its limiting distribution by  $\zeta_{pe}^{\det}$ . We conclude that:

$$e_{t+h}^m(\widehat{\Psi}|\mathfrak{S}_t) - e_{t+h}^m(\Psi|\mathfrak{S}_t) = i'_1 B^{-1} \left( \mathbf{e}^{-\delta C} - \mathbf{e}^{-\delta \widetilde{C}(\tau)} \right) \Lambda_{J_c}(\tau) + \zeta_{pe}^{\det}(\tau) + o_p \left( \sqrt{T} \right) \quad (48)$$

where we emphasize that the limiting distributions of the estimates depend on the time of estimation as a fraction of the sample size,  $\tau$ . Thus, this notation is general and encompasses the cases of recursive, split sample and rolling estimation.