Munich Personal RePEc Archive

"Go wild for a while!": A new asymptotically Normal test for forecast evaluation in nested models

Pincheira, Pablo and Hardy, Nicolás and Muñoz, Felipe

Universidad Adolfo Ibáñez, Universidad Finis Terrae

January 2021

Online at https://mpra.ub.uni-muenchen.de/105368/
MPRA Paper No. 105368, posted 27 Jan 2021 08:44 UTC

"Go wild for a while!": A new asymptotically Normal test for forecast evaluation in nested models[1]

Pablo Pincheira♣
School of Business
Universidad Adolfo Ibáñez

Nicolás Hardy
School of Economics and Business
Universidad Finis Terrae

Felipe Muñoz
School of Business
Universidad Adolfo Ibáñez

January 2021

Abstract

In this paper we present a new asymptotically normal test for out-of-sample evaluation in nested models. Our approach is a simple modification of a traditional encompassing test that is commonly known as Clark and West test (CW). The key point of our strategy is to introduce an independent random variable that prevents the traditional CW test from becoming degenerate under the null hypothesis of equal predictive ability. Using the approach developed by West (1996), we show that in our test the impact of parameter estimation uncertainty vanishes asymptotically. Using a variety of Monte Carlo simulations in iterated multi-step-ahead forecasts we evaluate our test and CW in terms of size and power. These simulations reveal that our approach is reasonably well-sized even at long horizons when CW may present severe size distortions. In terms of power, results are mixed but CW has an edge over our approach. Finally, we illustrate the use of our test with an empirical application in the context of the commodity currencies literature.

JEL Codes: C220, C530, E170, F370

Keywords: forecasting, random walk, out-of-sample, prediction, mean square prediction error.

---

[1] Corresponding Author: Pablo Pincheira Brown

♣ Pincheira: Diagonal Las Torres 2640, Peñalolén, Santiago, Chile. Email: pablo.pincheira@uai.cl.

†We thank Viviana Fernandez, Juan Pablo Medina and Augusto Castillo for valuable comments. We also thank participants of Universidad Finis Terrae workshop for interesting suggestions.

1. Introduction

In this paper we present a new asymptotically normal test for out-of-sample evaluation in the context of nested models. We label this test as "Wild Clark and West (WCW)." In essence, we propose a simple modification of the ENC-T (Clark and McCracken (2001) and Clark and West (2006, 2007)) core statistic that ensures asymptotic normality. The key point of our strategy is to introduce an independent random variable that prevents the CW test from becoming degenerate under the null hypothesis of equal predictive accuracy. Using West (1996) we show that "asymptotic irrelevance" applies, hence our test can ignore the effects of parameter uncertainty.

"*Mighty oaks from little acorns grow*." This is probably the best way to describe the forecast evaluation literature since the mid-90s. The seminal works of Diebold and Mariano (1995) and West (1996) (DMW) have flourished in many directions, attracting the attention of both scholars and practitioners in the quest for proper evaluation techniques. See West (2006), Clark and McCracken (2013a) and Giacomini and Rossi (2013) for great reviews on forecasting evaluation.

Considering forecasts as primitives, Diebold and Mariano (1995) show that under mild conditions on forecast errors and loss functions, standard time-series versions of the Central Limit Theorem apply, ensuring asymptotic normality for tests evaluating predictive performance. West (1996) considers the case in which forecasts are constructed with estimated econometric models. This is a critical difference with respect to Diebold and Mariano (1995) since forecasts are now polluted by estimation error.

Building on this insight, West (1996) develops a theory for testing population-level predictive ability (i.e., using estimated models to learn something about the true models). Two fundamental issues arise from West contribution: First, in some specific cases, parameter uncertainty is "asymptotically irrelevant," hence it is possible to proceed as proposed by Diebold and Mariano (1995). Second, although West's theory is quite general, it requires a full rank condition over the long-run variance of the objective function when parameters are set at their true values. A leading case in which this assumption is violated is in standard comparisons of Mean Squared Prediction Errors (MSPE) in nested environments.

As pointed out by West (2006): "*A rule of thumb is: if the rank of the data becomes degenerate when regression parameters are set at their population values, then a rank condition assumed in the previous sections likely is violated. When only two models are being compared, "degenerate" means identically zero*" West (2006) p.117. Clearly, in the context of two nested models, the null hypothesis of equal MSPE means that both models are exactly the same, which generates the violation of the rank condition in West (1996).

As nested models comparisons are extremely relevant in economics and finance, many efforts have been undertaken to deal with this issue. Some key contributions are those of Clark and McCracken (2001, 2005) and McCracken (2007), who use a different approach that allows for comparisons at the population level between nested models. Although in general, the derived asymptotic distributions are not standard, for some specific cases (e.g., no autocorrelation, conditional homoskedasticity of forecast errors, and one-step-ahead forecasts), the limiting distributions of the relevant statistics are free of nuisance parameters, and their critical values are provided in Clark and McCracken (2001).

While the contributions of many authors in the last 25 years have been important, our reading of the state of the art in forecast evaluation coincides with the view of Diebold (2015): *"[…] one must carefully tiptoe across a minefield of assumptions depending on the situation. Such assumptions include but are not limited to: 1) Nesting structure and nuisance parameters. Are the models nested, non-nested, or partially overlapping? 2) Functional form. Are the models linear or nonlinear? 3) Model disturbance properties. Are the disturbances Gaussian? Martingale differences? Something else? 4) Estimation sample. Is the pseudo-in-sample estimation period fixed? Recursively expanding? Something else? 5) Estimation method. Are the models estimated by OLS? MLE? GMM? Something else? And crucially: Does the loss function embedded in the estimation method match the loss function used for pseudo-out-of-sample forecast accuracy comparisons? 6) Asymptotics. What asymptotics are invoked?*" Diebold (2015) p. 3-4. Notably, the relevant limiting distribution generally depends on some of these assumptions.

In this context, there is a demand for straightforward tests that simplify the discussion in nested models comparisons. Of course, there are some attempts in the literature. For instance, one of the most used approaches in this direction is the test in Clark and West (2007). The authors show via simulations that standard normal critical values tend to work well with their test, even though, according to Clark and McCracken (2001), this statistic has a non-standard distribution. Moreover, when the null model is a martingale difference and parameters are estimated with rolling regressions, Clark and West (2006) show that their test is indeed asymptotically normal. Despite this and other particular cases, as stated in the conclusions of West (2006) review: "*One of the highest priorities for future work is the development of asymptotically normal or otherwise nuisance parameter-free tests for equal MSPE or mean absolute error in a pair of nested models. At present only special case results are available*". West (2006) p.131. Our paper addresses this issue.

Our WCW test can be viewed as a simple modification of the CW test. As noticed by West (1996), in the context of nested models comparisons, the CW core statistic becomes degenerate under the null hypothesis of equal predictive ability. Our suggestion is to introduce an independent random variable with a "small" variance in the core statistic. This random variable prevents our test from becoming degenerate under the null hypothesis, it keeps the asymptotic distribution centered around zero and eliminates the autocorrelation structure of the core statistic. While West (1996) asymptotic theory does not apply for CW (as it does not meet the full rank condition), it does apply for our test (as the variance of our test under the null hypothesis remains positive). In this sense, our approach not only prevents our test from becoming degenerate, but also ensures asymptotic normality relying on West (1996) results.

We also demonstrate that "asymptotic irrelevance" applies; hence the effects of parameter uncertainty can be ignored. As asymptotic normality and "asymptotic irrelevance" apply, our test is extremely friendly and easy to implement. Finally, one possible concern about our test is that it depends on one realization of one independent random variable. To partially overcome this issue, we also provide a smoothed version of our test that relies on multiple realizations of this random variable.

Most of the asymptotic theory for the CW test and other statistics developed in Clark and McCracken (2001, 2005) and McCracken (2007) focus almost exclusively on direct multi-step-ahead forecasts. However, with some exceptions (e.g., Clark and McCracken (2013b) and Pincheira and West (2016)), iterated multi-step-ahead forecasts have received much less attention. In part for this reason, we evaluate the performance of our test (relative to CW), focusing on iterated multi-step-ahead forecasts. Our simulations reveal that our approach is reasonably well-sized even at long horizons when CW may present severe size distortions. In terms of power, results are rather mixed although CW frequently exhibits some more power.

Finally, based on the commodity currencies literature, we provide an empirical illustration of our test. Following Chen, Rossi and Rogoff (2010,2011) and Pincheira and Hardy (2018, 2019a, 2019b), we evaluate the performance of the exchange rates of three major commodity producers economies (Australia, Chile, and South Africa) when predicting commodity prices. Consistent with previous literature, we find evidence of predictability for some of the commodities considered in this exercise.

The rest of this paper is organized as follows. Section 2 establishes the econometric setup, forecast evaluation framework and presents the WCW test. Section 3 demonstrates that the WCW is asymptotically normal, and that "asymptotic irrelevance" applies. Section 4 describes our DGPs and simulation setups. Section 5 discusses the simulation results. Section 6 provides an empirical illustration. Finally, section 7 concludes.

2. Econometric Setup

Consider the following two competing nested models for a target scalar variable y_{t+1}

$$y_{t+1} = X_t'\beta_1 + e_{1t+1} \quad \text{(model 1: null model)}$$
$$y_{t+1} = X_t'\beta_2 + Z_t'\gamma + e_{2t+1} \quad \text{(model 2: alternative model)}$$

Where e_{1t+1} and e_{2t+1} are both zero mean martingale difference processes, meaning that $E(e_{it+1}|F_t) = 0$ for $i = 1,2$ and F_t stands for the sigma-field generated by current and past values of X_t, Z_t and e_{it}. We will assume that e_{1t} and e_{2t} have finite and positive fourth moments.

The null hypothesis of interest is that $\gamma = 0$. This implies that $\beta_1 = \beta_2$ and $e_{1t+1} = e_{2t+1}$. This null hypothesis is also equivalent to equality in MSPE.

When the econometrician wants to test the null using an out-of-sample approach in this econometric context, Clark and McCracken (2001) derive the asymptotic distribution of a traditional encompassing statistic used, for instance, by Harvey, Leybourne and Newbold (1998)[2]. In essence, the ENC-t statistic proposed by Clark and McCracken (2001) studies the covariance between $\hat{e}_{1t+1}$ and $(\hat{e}_{1t+1} - \hat{e}_{2t+1})$. Accordingly, this test statistic takes the form:

$$ENC - t = \sqrt{P-1}\,\frac{P^{-1}\sum_{t=R+1}^{T} \hat{e}_{1t+1}(\hat{e}_{1t+1} - \hat{e}_{2t+1})}{\sqrt{\hat{\sigma}^2}}$$

Where $\hat{\sigma}^2$ is the usual variance estimator for $\hat{e}_{1t+1}(\hat{e}_{1t+1} - \hat{e}_{2t+1})$ and P is the number of out-of-sample forecasts under evaluation[3].

Even though West (1996) shows that the ENC-t is asymptotically Normal for non-nested models, this is not the case in nested environments. Note that one of the main assumptions in West (1996) theory is that the population counterpart of $\hat{\sigma}^2$ is strictly positive. This assumption is clearly violated when models are nested. To see this, recall that under the null of equal predictive ability, $\gamma = 0$ and $e_{1t+1} = e_{2t+1}$ for all t. In other words, the population prediction errors from both models are identical under the null and therefore $e_{1t+1}(e_{1t+1} - e_{2t+1})$ is exactly zero. Consequently $\sigma^2 = \mathbb{V}[e_{1t+1}(e_{1t+1} - e_{2t+1})] = 0$.

More precisely, notice that under the null:

$$e_{1t+1} = e_{2t+1}$$
$$e_{1t+1} - e_{2t+1} = 0$$

$$e_{1t+1}(e_{1t+1} - e_{2t+1}) = 0$$

$$\mathbb{E}[e_{1t+1}(e_{1t+1} - e_{2t+1})] = 0$$

---

[2] Other examples of encompassing tests include Chong and Hendry (1986) and Clements and Hendry (1993) to name a few.

[3] As pointed out by Clark and McCracken (2001), the HLN test is usually computed with regression-based methods. For this reason, we use $\sqrt{P-1}$ rather than $\sqrt{P}$.

$$\sigma^2 = \mathbb{V}[e_{1t+1}(e_{1t+1} - e_{2t+1})] = 0$$

It follows that the rank condition in West (1996) cannot be met as $\sigma^2 = 0$.

The main idea of our paper is to modify this ENC-t test to make it asymptotically Normal under the null. Our strategy requires the introduction of a sequence of independent random variables θ_t with variance ϕ^2 and expected value equal to 1. It is critical to notice that θ_t is not only i.i.d, but also independent from X_t, Z_t and e_{it}.

With this sequence in mind, we define our "Wild Clark and West" (WCW-t) statistic as

$$WCW - t = \sqrt{P-1}\ \frac{P^{-1}\sum_{t=R+1}^{T} \hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1})}{\sqrt{\widehat{S_{ff}}}}$$

Where $\widehat{S_{ff}}$ is a consistent estimate of the long-run variance of $\hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1})$ (e.g., Newey and West (1987,1994) or Andrews (1991), for instance).

In this case, under the null we have $e_{1t+1} = e_{2t+1}$, therefore:

$$\mathbb{E}[e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1})] = \mathbb{E}[e_{1t+1}(e_{1t+1} - \theta_t e_{1t+1})]$$

$$= \mathbb{E}[e_{1t+1}^2(1 - \theta_t)]$$

$$= \mathbb{E}[e_{1t+1}^2]\mathbb{E}(1 - \theta_t)$$

$$= \mathbb{E}[e_{1t+1}^2] * 0 \text{ (As we define } \mathbb{E}\theta_t = 1)$$

$$= 0 \text{ (hence our statistic is centered around 0)}$$

Besides, we have that under the null

$$\mathbb{V}[e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1})] = \mathbb{V}[e_{1t+1}(e_{1t+1} - \theta_t e_{1t+1})] = \mathbb{V}[e_{1t+1}^2(1 - \theta_t)]$$

$$= \mathbb{E}e_{1t+1}^4 * \mathbb{E}(1 - \theta_t)^2$$

$$= \phi^2 \mathbb{E}e_{1t+1}^4 > 0$$

The last result follows from the fact that $\mathbb{E}(1 - \theta_t)^2 = \mathbb{V}(\theta_t) = \phi^2$. Notice that this transformation is important: under the null hypothesis, even if $e_{1t+1}(e_{1t+1} - e_{2t+1})$ is identically zero for all t, the inclusion of θ_t prevents the core statistic from becoming degenerate, preserving a positive variance[4].

Additionally, under the alternative:

$$e_{1t+1} = y_{t+1} - X_t^{'}\beta_1$$

[4] It is also posible to show that the term $e_{1t+1}(e_{1t+1} - \theta_t e_{1t+1})$ has no autocorrelation under the null.

$$e_{2t+1} = y_{t+1} - X_t'\beta_2 - Z_t'\gamma$$

$$e_{2t+1} = e_{1t+1} - Z_t'\gamma - X_t'(\beta_2 - \beta_1)$$

Therefore:

$$\mathbb{E}[e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1})] = \mathbb{E}[e_{1t+1}(e_{1t+1} - \theta_t \left(e_{1t+1} - Z_t'\gamma - X_t'(\beta_2 - \beta_1)\right)]$$

$$= \mathbb{E}[e_{1t+1}^2(1-\theta_t)] + \mathbb{E}\left[e_{1t+1}\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)\theta_t\right]$$

$$= \mathbb{E}\left[e_{1t+1}\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)\right]$$

$$= \mathbb{E}\left[\left(X_t'\beta_2 + Z_t'\gamma + e_{2t+1} - X_t'\beta_1\right)\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)\right]$$

$$= \mathbb{E}\left[\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1) + e_{2t+1}\right)\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)\right]$$

$$= \mathbb{E}\left[\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)^2\right] + \mathbb{E}\left[e_{2t+1}\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)\right]$$

$$= \mathbb{E}\left[\left(Z_t'\gamma + X_t'(\beta_2 - \beta_1)\right)^2\right] > 0$$

And consequently, our test is one-sided.

Finally, there are two possible concerns with the implementation of our WCW-t statistic. The first one is about the choice of $\mathbb{V}(\theta_t) = \phi^2$. Even though this decision is arbitrary, we give the following recommendation: ϕ^2 should be "small"; the idea of our test is to recover asymptotic normality under the null hypothesis, something that could be achieved for any value of $\phi^2 > 0$. However, if ϕ^2 is "too big," it may simply erode the predictive content under the alternative hypothesis, deteriorating the power of our test. Notice that a "small" variance for some DGPs could be a "big" one for others, for this reason, we propose to take ϕ as a small percentage of the sample counterpart of $\sqrt{\mathbb{V}(e_{2t+1})}$. As we discuss later in Section 4, we consider three different standard deviations with reasonable size and power results: $\phi = \{0.01 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}; 0.02 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}; 0.04 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}\}$ (1 percent, 2 percent and 4 percent of the standard deviation of $\hat{e}_{2t+1}$). We emphasize that $\mathbb{V}(\hat{e}_{2t+1})$ is the sample variance of the estimated forecast errors. Obviously, our test tends to be better sized as ϕ grows, at the cost of some power.

Second, notice that our test depends on K=1 realization of the sequence θ_t. One reasonable concern is that this randomness could strongly affect our WCW-t statistic (even for "small" values of the ϕ^2 parameter). In other words, we would like to avoid significant changes in our statistic generated by the randomness of θ_t. Additionally, as we report in Section 4, our simulations suggest that using just one realization of the sequence θ_t sometimes may significantly reduce the power of our test relative to CW. To tackle both issues, we propose to smooth the randomness of our approach by considering K different WCW-t statistics constructed with different and independent sequences of θ_t. Our proposed test is the simple average of these K standard normal WCW-t statistics, adjusted by the correct variance of the average as follows:

$$WCW(K) - t = \frac{\sum_{k=1}^{K} WCW_k}{\sqrt{\sum_{j=1}^{K}\sum_{i=1}^{K}\rho_{i,j}}} \qquad (1)$$

Where WCW_k is the k-th realization of our statistic and $\rho_{i,j}$ is the sample correlation between the i-th and j-th realization of the WCW-t statistics. Interestingly, as we discuss in Section 4, when using K=2 the size of our test is usually stable, but it significantly improves the power of our test.

3. Asymptotic Normality

Since most of our results rely on West (1996), here we introduce some of his results and notation. For clarity of exposition, we focus on one-step-ahead forecasts. The generalization to multi-step-ahead forecasts is cumbersome in notation but straightforward.

Let $f_{t+1} = e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1}) = (Y_{t+1} - X_t'\beta_1^*)(Y_{t+1} - X_t'\beta_1^* - \theta_t[Y_{t+1} - X_t'\beta_2^* - Z_t'\gamma^*])$ be our loss function. We use "*" to emphasize that f_t depends on the true population parameters, hence $f_{t+1} \equiv f_{t+1}(\beta^*)$ where $\beta^* = [\beta_1^*, \beta_2^*, \gamma^*]'$. Additionally, let $\hat{f}_{t+1} \equiv f_{t+1}(\hat{\beta}_t) = \hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1}) = (Y_{t+1} - X_t'\hat{\beta}_{1t})(Y_{t+1} - X_t'\hat{\beta}_{1t} - \theta_t[Y_{t+1} - X_t'\hat{\beta}_{2t} - Z_t'\hat{\gamma}_t])$ be the sample counterpart of f_{t+1}. Notice that $f_{t+1}(\hat{\beta}_t)$ rely on estimates of β^*, and as a consequence, $f_{t+1}(\hat{\beta}_t)$ is polluted by estimation error. Moreover, notice the subindex in $\hat{\beta}_t$: the out-of-sample forecast errors ($\hat{e}_{1t+1}$ and $\hat{e}_{2t+1}$) depends on the estimates $\hat{\beta}_t$ constructed with the relevant information available up to time t. These estimates can be constructed using either rolling, recursive, or fixed windows. See West (1996, 2006) and Clark and McCracken (2013a) for more details about out-of-sample evaluations.

Let $\mathbb{E}f_t = \mathbb{E}[e_{1t}(e_{1t} - \theta_t e_{2t})]$ the expected value of our loss function. As considered in Diebold and Mariano (1995), if predictions do not depend on estimated parameters, then under weak conditions, we can apply the central limit theorem:

$$\sqrt{P}\left(P^{-1}\sum_t f_{t+1} - \mathbb{E}f_t\right) \sim_A N\left(0, S_{ff}\right) \qquad (2)$$

$$S_{ff} \equiv \sum_{j=-\infty}^{\infty} \mathbb{E}\{(f_{t+1} - \mathbb{E}f_t)\left(f_{t+1-j} - \mathbb{E}f_{t+1-j}\right)\}$$

Where $S_{ff} > 0$ stands for the long-run variance of the scalar f_{t+1}. However, one key technical contribution in West (1996) is to notice that when forecasts are constructed with estimated rather than true, unknown, population parameters, some terms in expression (2) must be adjusted. We remark here that we observe $\hat{f}_{t+1} = \hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1})$ rather than $f_{t+1} = e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1})$. To see how parameter uncertainty may play an important role, under assumptions A.1-A.4 in the Appendix, West (1996) shows that a second-order expansion of $f_t(\hat{\beta})$ around β yields

$$P^{-\frac{1}{2}} \sum_{t=R}^{T-1} \left(\hat{f}_{t+1} - \mathbb{E}f_t\right) = P^{-\frac{1}{2}} \sum_{t=R}^{T-1} (f_{t+1} - \mathbb{E}f_t) + F\left(\frac{P}{R}\right)^{\frac{1}{2}} \left(BR^{\frac{1}{2}}\overline{H}\right) + o_p(1) \quad (3)$$

Where $F = \frac{\partial \mathbb{E} f_t(\beta^*)}{\partial \beta}$, R denotes the length of the initial estimation window, T is the total sample size (T=R+P), while B and $\overline{H}$ will be defined shortly.

Recall that in our case, under the null hypothesis $\mathbb{E}f_{t+1} = \mathbb{E}[e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1})] = 0$, hence expression (3) is equivalent to

$$P^{-\frac{1}{2}} \sum_{t=R}^{T-1} \hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1}) = P^{-\frac{1}{2}} \sum_{t=R}^{T-1} e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1}) + F\left(\frac{P}{R}\right)^{\frac{1}{2}} \left(BR^{\frac{1}{2}}\overline{H}\right) + o_p(1)$$

Note that according to West (2006) pp.112, and in line with Assumption 2 in West (1996) pp.1070-1071, the estimator of the regression parameters satisfies

$$\widehat{\beta_t} - \beta^* = B(t)H(t),$$

Where $B(t)$ is kxq, $H(t)$ is $qx1$ with

a) $B(t) \overset{a.s}{\rightarrow} B, B$ a matrix of rank k;
b) $H(t)= t^{-1} \sum_{s=1}^{t} h_s(\beta^*)$ if the estimation method is recursive, $H(t)= R^{-1} \sum_{s=t-R+1}^{t} h_s(\beta^*)$ if it is rolling or $H(t)= R^{-1} \sum_{s=1}^{R} h_s(\beta^*)$ if it is fixed. $h_s(\beta^*)$ is a $qx1$ orthogonality condition that satisfies. Notice that $\overline{H} = P^{-1} \sum_{t=R}^{T-1} H(t)$.
c) $Eh_s(\beta^*) = 0$.

As explained in West (2006): *"Here, h_t can be considered as the score if the estimation method is ML, or the GMM orthogonality condition if GMM is the estimator. The matrix $B(t)$ is the inverse of the Hessian if the estimation method is ML or a linear combination of orthogonality conditions when using GMM, with large sample counterparts B."* West (2006) pp.112.

Notice that Eq.(3) clearly illustrates the point: $P^{-\frac{1}{2}} \sum_t \hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t \hat{e}_{2t+1})$ can be decomposed into two parts. The first term of the RHS is the population counterpart, whereas the second term captures the sequence of estimates of β^* (in other words, terms arising because of parameter uncertainty). Then, as $P, R \to \infty$, we can apply the expansion in West (1996) as long as assumptions A1-A4 holds. The key point is that a proper estimation of the variance in Eq.(3) must account for: i) The variance of the first term of the RHS ($S_{ff} = \phi^2 \mathbb{E} e_{1t+1}^4 > 0$, i.e., the variance when there is no uncertainty about the population parameters), ii) The variance of the second term of the RHS, associated with parameter uncertainty, and iii) the covariance between both terms. Notice, however, that parameter uncertainty may be "asymptotically irrelevant" (hence ii) and iii) may be ignored) in the following cases: 1) $\frac{P}{R} \to 0$ as $P, R \to \infty$, 2) A fortunate cancellation between ii) and iii) or 3) $F = 0$.

In our case:

$$F = \mathbb{E}\frac{\partial f_t(\beta)}{\partial \beta}|_{\beta=\beta^*} = [\mathbb{E}\frac{\partial f_t(\beta)}{\partial \beta_1}|_{\beta=\beta^*}, \mathbb{E}\frac{\partial f_t(\beta)}{\partial \beta_2}|_{\beta=\beta^*}, \mathbb{E}\frac{\partial f_t(\beta)}{\partial \gamma}|_{\beta=\beta^*}]$$

Where

$$f_t(\beta) = (Y_{t+1} - X_t'\beta_1)(Y_{t+1} - X_t'\beta_1 - \theta_t[Y_{t+1} - X_t'\beta_2 - Z_t'\gamma])$$

$$f_t(\beta) = (Y_{t+1} - X_t'\beta_1)^2 - (Y_{t+1} - X_t'\beta_1)\theta_t(Y_{t+1} - X_t'\beta_2 - Z_t'\gamma)$$

$$\frac{\partial f_t(\beta)}{\partial \beta_1} = -2(Y_{t+1} - X_t'\beta_1)X_t + \theta_t(Y_{t+1} - X_t'\beta_2 - Z_t'\gamma)X_t$$

Note that under the null, $\gamma^* = 0, \beta_1^* = \beta_2^*$ and recall that $\mathbb{E}\theta_t = 1$, therefore

$$\mathbb{E}\frac{\partial f_t(\beta)}{\partial \beta_1}|_{\beta=\beta^*} = -2\mathbb{E}e_{1t+1}X_t + \mathbb{E}\theta_t\mathbb{E}e_{1t+1}X_t = 0$$

With a similar argument, it is easy to show that

$$\mathbb{E}\frac{\partial f_t(\beta)}{\partial \beta_2}|_{\beta=\beta^*} = \mathbb{E}X_t'e_{1t+1}\mathbb{E}\theta_t = 0$$

Finally

$$\mathbb{E}\frac{\partial f_t(\beta)}{\partial \gamma} = (Y_{t+1} - X_t'\beta_1)\theta_t Z_t \Rightarrow \mathbb{E}\frac{\partial f_t(\beta)}{\partial \gamma}|_{\beta=\beta^*} = \mathbb{E}\theta_t\mathbb{E}Z_t'e_{1t+1} = \mathbb{E}Z_t'e_{1t+1} = 0$$ [5]

Hence, in our case "asymptotic irrelevance" applies as $F = 0$ and Eq. (3) reduces simply to

$$P^{-\frac{1}{2}}\sum_{t=R}^{T-1}\hat{e}_{1t+1}(\hat{e}_{1t+1} - \theta_t\hat{e}_{2t+1}) = P^{-\frac{1}{2}}\sum_{t=R}^{T-1}e_{1t+1}(e_{1t+1} - \theta_t e_{2t+1}) + o_p(1)$$

[5] This result follows from the fact that we define e_{1t+1} as a martingale difference respect to X_t and Z_t.

In other words, we could simply replace true errors by estimated out-of-sample errors and forget about parameter uncertainty, at least asymptotically.

4. Monte Carlo simulations

We consider three different DGPs for our simulations. To save space, we only report here results for recursive windows, although in general terms, results with rolling windows are similar and they are available upon request. For large sample exercises we consider an initial estimation window of R=450 and a prediction window of P=450 (T=900), while for small sample exercises, we consider R=90 and P=90 (T=180).

For each DGP, we run 2,000 independent replications. We evaluate the CW test and our test computing iterated multi-step-ahead forecasts at several forecasting horizons from h=1 up to h=30. As discussed at the end of Section 2, we compute our test using K=1 and K=2 realizations of our WCW-t statistic. Additionally, for each simulation, we consider three different standard deviations of θ_t: $\phi = \{0.01 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}; 0.02 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}; 0.04 * \sqrt{\mathbb{V}(\hat{e}_{2t+1})}\}$ (1 percent, 2 percent and 4 percent of the standard deviation of $\hat{e}_{2t+1}$) . We emphasize that $\mathbb{V}(\hat{e}_{2t+1})$ is the sample variance of the out-of-sample forecast errors, and it is calculated for each simulation.

Finally, we evaluate the usefulness of our approach using the iterated multistep ahead method for the three DGPs under evaluation[6]. We report our results comparing the CW and the WCW-t test using one-sided standard normal critical values at the 10% and 5% significance level (a summary of the results considering a 5% significance level can be found in the Appendix section). For simplicity, in each simulation we consider only homoscedastic, i.i.d normally distributed shocks.

4.1 DGP 1

Our first DGP assumes a white noise for the null model. We consider a case like this given its relevance in finance and macroeconomics. Our setup is very similar to simulation experiments in Pincheira and West (2006), Stambaugh (1999), Nelson and Kim (1993), and Mankiw and Shapiro (1986).

Null Model:

$$Y_{t+1} = \varepsilon_{t+1}$$

Alternative Model:

[6] Notice that the iterated method uses an auxiliary equation for the construction of the multistep ahead forecasts. Here we stretch the argument of "asymptotic irrelevance" and we assume that parameter uncertainty on the auxiliary equation plays no role.

$$Y_{t+1} = \alpha_y + \gamma r_t + \varepsilon_{t+1}$$

$$r_{t+1} = \alpha_r + \rho_1 r_t + \rho_2 r_{t-1} + \ldots + \rho_p r_{t-p} + v_{t+1}$$

We set our parameters as follows

$$\alpha_y = \alpha_r = \rho_3 = \cdots = \rho_P = 0$$

$$\mathbb{V}(\varepsilon_{t+1}) = \sigma_\varepsilon^2$$

$$\mathbb{V}(v_{t+1}) = \sigma_v^2$$

$$Corr(\varepsilon_{t+1}, v_{t+1}) = \psi$$

ρ_1	ρ_2	σ_ε^2	σ_v^2	ψ	γ under H_0	γ under H_A
1.19	–0.25	$(1.75)^2$	$(0.075)^2$	0	0	-2

The null hypothesis posits that Y_{t+1} follows a *no-change* martingale difference. Additionally, the alternative forecast for multi-step-ahead horizons is constructed iteratively through an AR(p) on r_{t+1}. This is the same parametrization considered in Pincheira and West (2016), and it is based on a monthly exchange rate application in Clark and West (2006). Therefore, Y_{t+1} represents the monthly return of a U.S dollar bilateral exchange rate and r_t is the corresponding interest rate differential.

4.2 DGP 2

Our second DGP is mainly inspired in macroeconomic data, and it is also considered in Pincheira and West (2016) and Clark and West (2007). This DGP is based on models exploring the relationship between U.S GDP growth and the Federal Reserve Bank of Chicago's factor index of economic activity.

Null Model:

$$Y_{t+1} = \alpha_y + \delta r_t + \varepsilon_{t+1}$$

Alternative Model:

$$Y_{t+1} = \alpha_y + \delta Y_t + \gamma_1 r_t + \gamma_2 r_t + \cdots + \gamma_p r_{t-p} + \varepsilon_{t+1}$$

$$r_{t+1} = \alpha_r + 0.804 r_t - 0.221 r_{t-1} + 0.226 r_{t-2} - 0.205 r_{t-3} + v_{t+1}$$

We set our parameters as follows

$$\alpha_y = 2.237$$

$$\alpha_r = \gamma_5 = \cdots = \gamma_p = 0$$

$$\mathbb{V}(\varepsilon_{t+1}) = \sigma_\varepsilon^2$$

$$\mathbb{V}(v_{t+1}) = \sigma_v^2$$

$$Corr(\varepsilon_{t+1}, v_{t+1}) = \psi$$

γ_1 under H_0	γ_2 under H_0	γ_3 under H_0	γ_4 under H_0
0	0	0	0

γ_1 under H_A	γ_2 under H_A	γ_3 under H_A	γ_4 under H_A
3.363	-0.633	-0.377	-0.529
δ	σ_ε^2	σ_v^2	ψ
0.261	10.505	0.366	0.528

4.3 DGP 3

Our last DGP follows Busetti and Marcucci (2013) and considers a very simple VAR(1) process:

Null Model:

$$Y_{t+1} = \mu_y + \phi_y Y_t + \varepsilon_{t+1}$$

Alternative Model:

$$Y_{t+1} = \mu_y + \phi_y Y_t + cX_t + \varepsilon_{t+1}$$

$$X_{t+1} = \mu_x + \phi_x X_t + v_{t+1}$$

We set our parameters as follows

$$\mu_y = \mu_x = 0$$

$$\mathbb{V}(\varepsilon_{t+1}) = \sigma_\varepsilon^2$$

$$\mathbb{V}(v_{t+1}) = \sigma_v^2$$

$$Corr(\varepsilon_{t+1}, v_{t+1}) = \psi$$

ϕ_y	ϕ_x	σ_ε^2	σ_v^2	ψ	c under H_0	c under H_A
0.8	0.8	1	1	0	0	0.5

5 Simulation Results

This section reports exclusively results for a nominal size of 10%. To save space, we consider only results with a recursive scheme. Results with rolling windows are similar, and they are available upon request[7]. For each simulation, we consider θ_t i.i.d normally distributed with mean one and variance ϕ^2. Tables 1-6 show results on size considering different choices for $\mathbb{V}(\theta_t) = \phi^2$ and K, as suggested at the end of Section 2. The last row of each table reports the average size for each test across the 30 forecasting horizons. Tables 7-12 are akin to Tables 1-6, but they report results on power. Likewise to Tables 1-6, the last row of each table reports the average power for each test across the 30 forecasting horizons. Our analysis with a nominal size of 5% carries the same message. A summary of these results can be found in the Appendix.

5.1 Simulation Results: Size

Table 1 reports results for the case of a martingale sequence (i.e DGP1) using large samples (P=R=450 and T=900). From the second column of Table 1, we observe that the CW test is modestly undersized. The empirical size of nominal 10% tests ranges from 6% to 8%, with an average size across the 30 forecasting horizons of 6%. These results are not surprising. For instance, for the case of a martingale sequence, Clark and West (2006) comment that: "*our statistic is slightly undersized, with actual sizes ranging from 6.3% […] to 8.5%*" Clark and West (2006), pp. 172-173. Moreover, Pincheira and West (2016), using iterated multi-step ahead forecasts, find very similar results.

Our test seems to behave reasonably well. Across the nine different exercises presented in Table 1, the empirical size of our WCW test ranges from 7% to 11%. Moreover, the last row indicates that the average size of our exercises ranges from 0.08 ($\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$) to 0.10 (e.g., all exercises considering $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$). Notably, our results using "the highest variance" $0.04 * \sigma(e_2)$ range from 9% to 11%, with an average size of 10% in the two cases. As we discuss in the following section, in some cases, this outstanding result comes at the cost of some reduction in power.

Table 2 is akin to Table 1, but considering simulations with small samples (P=R=90 and T=180). While the overall message is very similar, the CW test behaves remarkably well, with an empirical size ranging from 8% to 10% and an average size of 9%. Additionally, our test also shows a good size behavior, but with mild distortions in some experiments.

[7] Results of the recursive method are more interesting to us for the following reason: For DGP1, Clark and West (2006) show that the CW statistic with rolling windows is indeed asymptotically normal. In this regard, the recursive method may be more interesting to discuss due to the expected departure from normality in the CW test.

Despite these cases, in 6 out of 9 exercises our test displays an average size of 10% across different forecast horizons. The main message of Tables 1-2 is that our test behaves reasonably well, although there are no great improvements (nor losses) compared to CW.

Table 1: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP1 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.07	0.08	0.08	0.09	0.09	0.09	0.10
2	0.07	0.09	0.08	0.10	0.09	0.09	0.11
3	0.08	0.08	0.07	0.10	0.09	0.09	0.11
6	0.07	0.09	0.08	0.10	0.09	0.10	0.11
12	0.06	0.09	0.07	0.10	0.09	0.09	0.11
15	0.06	0.08	0.08	0.09	0.09	0.11	0.10
18	0.06	0.09	0.08	0.10	0.09	0.09	0.10
21	0.06	0.08	0.08	0.10	0.10	0.09	0.11
24	0.06	0.08	0.08	0.11	0.08	0.10	0.11
27	0.06	0.08	0.08	0.10	0.10	0.09	0.10
30	0.07	0.09	0.07	0.11	0.10	0.10	0.11
Average Size	0.06	0.08	0.08	0.10	0.09	0.10	0.10

Notes: Table 1 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\widehat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 2: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP1 and a small sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.08	0.08	0.08	0.09	0.08	0.10	0.09
2	0.09	0.09	0.09	0.09	0.09	0.10	0.11
3	0.10	0.09	0.09	0.09	0.10	0.11	0.11
6	0.09	0.09	0.09	0.11	0.10	0.11	0.11
12	0.09	0.09	0.09	0.10	0.10	0.11	0.11
15	0.09	0.10	0.10	0.11	0.10	0.12	0.11
18	0.09	0.09	0.11	0.10	0.12	0.11	0.12
21	0.10	0.10	0.10	0.11	0.10	0.11	0.11
24	0.10	0.11	0.10	0.12	0.11	0.12	0.11
27	0.09	0.10	0.10	0.11	0.11	0.12	0.11
30	0.10	0.10	0.11	0.10	0.11	0.10	0.12
Average Size	0.09	0.10	0.10	0.11	0.10	0.11	0.11

Notes: Table 2 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=180 (R=90 and P=90). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 3 reports our results for DGP2 using large samples (P=R=450 and T=900). In this case, the empirical size of the CW test ranges from 8% to 16%, with an average size of 13%. Notably, the CW test is undersized at "short" forecasting horizons ($h \leq 3$) and oversized at long forecasting horizons ($h \geq 12$). This is consistent with the results reported in Pincheira and West (2016) for the same DGP using a rolling scheme: *"[…] the CW test has a size ranging from 7% to 13%. It tends to be undersized at shorter horizons ($h \leq 3$), oversized at longer horizons ($h \geq 6$)."* Pincheira and West (2013), pp. 313.

In contrast, our test tends to be considerably better sized. Across all exercises, the empirical size of the WCW ranges from 8% to 12%. Moreover, the average size for each one of our tests is in the range of 10% to 11%. In sharp contrast with CW, our test has a "stable" size and does not become increasingly oversized with the forecasting horizon. In specific, for h=30, the empirical size of our test across all exercises is exactly 10%, while CW has an empirical size of 15%. In this sense, our test offers better protection to the null hypothesis at long forecasting horizons.

Table 4 is akin to Table 3, but considering a smaller sample. The overall message is similar; however, both CW and our test become oversized. Despite these size distortions in both tests, we emphasize that our test performs comparatively better relative to CW in almost every exercise. For instance, using a standard deviation of $\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$ or $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$, our test is reasonably well-sized across all exercises. The worst results are found for $\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$; however, our worst exercise, with K=2, is still better (or equal) sized than CW for all horizons. The intuition of $\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$ presenting the worst results is in fact by construction; recall that for $\sigma(\theta_t) = 0$, our test coincides with CW, hence, as the variance of θ_t becomes smaller, it is likely to expect stronger similarities between CW and our test. In a nutshell, Tables 3-4 indicate that our test is reasonably well sized, with some clear benefits compared to CW for long horizons (e.g., $h \geq 12$), as CW becomes increasingly oversized.

Table 3: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP2 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.08	0.09	0.09	0.11	0.11	0.11	0.11
2	0.08	0.08	0.10	0.09	0.11	0.09	0.12
3	0.08	0.09	0.08	0.10	0.09	0.10	0.11
6	0.09	0.10	0.09	0.11	0.09	0.11	0.09
12	0.13	0.10	0.11	0.10	0.11	0.10	0.11
15	0.15	0.11	0.10	0.11	0.10	0.12	0.10
18	0.15	0.11	0.12	0.11	0.12	0.11	0.12
21	0.16	0.10	0.10	0.10	0.10	0.10	0.10
24	0.15	0.11	0.11	0.11	0.11	0.11	0.11
27	0.15	0.12	0.10	0.12	0.10	0.12	0.10
30	0.15	0.10	0.10	0.10	0.10	0.10	0.10
Average Size	0.13	0.10	0.10	0.10	0.10	0.10	0.11

Notes: Table 3 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\widehat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 4: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP2 and a small sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.09	0.10	0.09	0.10	0.10	0.11	0.10
2	0.10	0.10	0.10	0.10	0.11	0.11	0.11
3	0.10	0.10	0.10	0.12	0.11	0.12	0.12
6	0.11	0.10	0.11	0.10	0.11	0.11	0.11
12	0.14	0.11	0.14	0.10	0.14	0.10	0.14
15	0.15	0.13	0.13	0.12	0.11	0.11	0.11
18	0.15	0.13	0.13	0.12	0.12	0.11	0.11
21	0.15	0.12	0.13	0.11	0.11	0.11	0.11
24	0.17	0.13	0.14	0.11	0.13	0.10	0.12
27	0.16	0.12	0.14	0.11	0.11	0.11	0.11
30	0.17	0.14	0.14	0.12	0.13	0.12	0.12
Average Size	0.14	0.12	0.13	0.11	0.12	0.11	0.11

Notes: Table 4 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=180 (R=90 and P=90). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Finally, Tables 5-6 show our results for DGP3 using large samples (P=R=450 and T=900) and small samples (P=R=90 and T=180), respectively. The main message is very similar to that obtained from DGP2: CW is slightly undersized at short forecasting horizons (e.g., $h \leq 3$); and increasingly oversized at longer horizons ($h \geq 12$). In contrast, our test either does not exhibit this pattern with the forecasting horizon or, when it does, it is milder. Notably, for long horizons (e.g., h=30) our test is always better sized than CW. As in the previous DGP, our test works very well using "the higher variance" $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$, and becomes increasingly oversized as the standard deviation approaches to zero. Importantly, using the two highest variances ($\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$ and $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$) our worst results are empirical sizes of 16%; in sharp contrast, the worst entries for CW are 20% and 22%.

All in all, Tables 1 through 6 provide a similar message: On average, our test seems to be better sized, specially at longer forecasting horizons. The size of our test improves with a higher $\sigma(\theta_t)$, but as we will see in the following section, sometimes this improvement comes at the cost of a slightly reduction in power.

Table 5: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP3 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.08	0.08	0.08	0.09	0.09	0.10	0.10
2	0.08	0.09	0.09	0.11	0.10	0.10	0.11
3	0.09	0.09	0.10	0.10	0.11	0.12	0.12
6	0.11	0.12	0.13	0.12	0.13	0.12	0.13
12	0.13	0.12	0.14	0.12	0.14	0.11	0.12
15	0.14	0.11	0.14	0.11	0.12	0.10	0.11
18	0.15	0.12	0.13	0.11	0.11	0.11	0.11
21	0.14	0.12	0.12	0.12	0.11	0.12	0.11
24	0.12	0.11	0.11	0.11	0.11	0.11	0.10
27	0.15	0.10	0.12	0.11	0.12	0.10	0.11
30	0.15	0.10	0.10	0.10	0.10	0.10	0.10
Average Size	0.13	0.11	0.12	0.11	0.12	0.11	0.11

Note: Table 5 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size results across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposal using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 6: Empirical size comparisons between CW and WCW tests with nominal size of 10%, considering DGP3 and a small sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.07	0.07	0.07	0.07	0.07	0.08	0.08
2	0.09	0.09	0.09	0.09	0.09	0.09	0.10
3	0.10	0.11	0.10	0.12	0.11	0.11	0.12
6	0.13	0.13	0.14	0.14	0.14	0.14	0.14
12	0.16	0.15	0.17	0.13	0.16	0.13	0.16
15	0.17	0.16	0.18	0.14	0.16	0.13	0.16
18	0.17	0.14	0.17	0.13	0.15	0.12	0.14
21	0.18	0.15	0.17	0.13	0.14	0.11	0.12
24	0.19	0.16	0.16	0.14	0.14	0.13	0.12
27	0.20	0.15	0.17	0.13	0.14	0.12	0.12
30	0.22	0.17	0.19	0.15	0.16	0.13	0.14
Average Size	0.16	0.14	0.16	0.13	0.14	0.12	0.14

Notes: Table 6 presents empirical sizes for the CW test and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average size results across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2

($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=180 (R=90 and P=90). We evaluate the CW test and our proposal using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

5.2 Simulation Results: Power

The intuition of our test is that we achieve normality introducing a random variable that prevents the core statistic of the CW test from becoming degenerate under the null hypothesis. As reported in the previous section, our test sometimes displays better size relative to CW, especially at long forecasting horizons. However, the presence of this random variable may also erode some of the predictive content of model 2, and consequently, it may also erode the power of our test. As we will see in this section, results in terms of power are mixed: sometimes CW exhibits superior power, sometimes the differences are negligible, and sometimes WCW displays higher power (although this is less frequent).

Tables 7 and 8 report power results for DGP1 considering large and small samples, respectively. Table 7 shows results that are, more or less, consistent with the previous intuition: the worst results are found for the highest standard deviation ($\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$) and one sequence of realizations of θ_t (K=1). In this sense, the good results in terms of size reported in the previous section come at the cost of a slight reduction in power. In this case, the average loss of power across the 30 forecasting horizons is about 6% (55% for CW and 49% for our "less powerful" exercise). Notice, however, that averaging two independent realizations of our test (e.g., K=2) or reducing $\sigma(\theta_t)$, rapidly enhance the power of our test. Actually, with $K = 2$ and a low variance of $\sigma(\theta_t)$ the power of our test becomes very close to CW. The best results in terms of power are found for the smallest variance. This can be partially explained for the fact that the core statistic of our test becomes exactly the CW core statistic when the variance (θ_t) approaches zero. Table 8 shows results mostly in the same line; although this time figures are much lower due to the small sample. Importantly, differences in terms of power are almost negligible between our approach and CW.

Table 7: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP1 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	0.99	1.00
6	0.97	0.97	0.97	0.96	0.97	0.93	0.95
12	0.69	0.68	0.68	0.63	0.64	0.55	0.60
15	0.50	0.49	0.50	0.45	0.48	0.38	0.43
18	0.37	0.36	0.36	0.33	0.35	0.28	0.30
21	0.26	0.26	0.27	0.24	0.26	0.21	0.25
24	0.21	0.21	0.21	0.20	0.22	0.19	0.20
27	0.18	0.18	0.19	0.18	0.19	0.16	0.18
30	0.17	0.17	0.17	0.16	0.17	0.15	0.16
Average Power	0.55	0.54	0.55	0.52	0.54	0.49	0.52

Notes: Table 7 presents power results for CW and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average power across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 8: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP1 and a small sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	0.78	0.78	0.78	0.77	0.78	0.73	0.77
2	0.77	0.76	0.77	0.74	0.76	0.70	0.74
3	0.71	0.71	0.71	0.70	0.70	0.66	0.68
6	0.51	0.50	0.50	0.49	0.50	0.46	0.49
12	0.27	0.27	0.28	0.26	0.28	0.25	0.28
15	0.23	0.23	0.23	0.23	0.23	0.22	0.23
18	0.21	0.21	0.21	0.20	0.21	0.20	0.21
21	0.19	0.19	0.19	0.19	0.20	0.18	0.19
24	0.18	0.17	0.17	0.17	0.18	0.17	0.19
27	0.17	0.17	0.18	0.17	0.19	0.16	0.18
30	0.17	0.17	0.17	0.17	0.17	0.17	0.17
Average Power	0.32	0.32	0.32	0.31	0.32	0.30	0.32

Notes: Same notes as in Table 7. The only difference is that in Table 8 the sample size is T=180 (R=90 and P=90).

Tables 9 and 10 report power results for DGP2, considering large and small samples, respectively. Contrary to DGP1, now power reductions using our approach are important for some exercises. For instance, in Table 10, CW has 20% more rejections than our "less powerful" exercise. In this sense, asymptotic normality and good results for $\sigma(\theta_t) = 0.04 * \sigma(e_2)$ in terms of size, comes along with an important reduction in power. As noticed before, the power of our test rapidly improves with K>1 or with a smaller $\sigma(\theta_t)$. For instance, in Table 10, for the case of $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$, if we consider K=2 instead of K=1, the average power improves from 37% to 43%. Moreover, if we keep K=2 and reduce $\sigma(\theta_t)$ to $\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$, differences in power compared to CW are small.

Table 9: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP2 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	0.98	1.00
12	0.86	0.73	0.81	0.49	0.62	0.32	0.42
15	0.61	0.40	0.51	0.28	0.33	0.20	0.23
18	0.48	0.28	0.39	0.19	0.25	0.15	0.18
21	0.41	0.22	0.29	0.17	0.20	0.14	0.16
24	0.36	0.17	0.23	0.14	0.15	0.13	0.14
27	0.31	0.16	0.19	0.12	0.14	0.10	0.13
30	0.29	0.14	0.16	0.13	0.12	0.12	0.11
Average Power	0.65	0.54	0.59	0.47	0.51	0.42	0.46

Notes: Table 9 presents power results for CW and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports average power results across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 10: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP2 and a small sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.99	0.97	0.97	0.91	0.95	0.80	0.90
6	0.94	0.91	0.93	0.84	0.90	0.71	0.83
12	0.58	0.50	0.55	0.38	0.45	0.28	0.35
15	0.46	0.38	0.45	0.29	0.35	0.23	0.27
18	0.41	0.33	0.38	0.26	0.29	0.20	0.22
21	0.38	0.30	0.36	0.21	0.28	0.17	0.22
24	0.36	0.28	0.33	0.21	0.25	0.17	0.20
27	0.36	0.27	0.32	0.20	0.25	0.16	0.19
30	0.36	0.28	0.33	0.20	0.24	0.16	0.19
Average Power	0.58	0.52	0.57	0.44	0.50	0.37	0.43

Notes: Same notes as in Table 9. The only difference is that in Table 10 the sample size is T=180 (R=90 and P=90).

Finally, Tables 11 and 12 report power results for DGP3, considering large and small samples, respectively. In most cases reductions in power are small (if any). For instance, our "less powerful exercise" in Table 11 has an average power only 3% below CW (although there are some important differences at long forecasting horizons such as h=30). However, as commented previously, the power of our test rapidly improves when considering $K = 2$; in this case, differences in power are fairly small for all exercises. Notably, in some cases we find tiny (although consistent) improvements in power over CW; for instance, using the smallest standard deviation and K=2, our test is “as powerful” as CW, and sometimes even slightly more powerful for longer horizons (e.g., h>18).

Table 11: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP3 and a large sample.

Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	0.98	0.98	0.98	0.98	0.98	0.97	0.98
15	0.90	0.90	0.90	0.90	0.91	0.88	0.90
18	0.82	0.82	0.82	0.80	0.82	0.77	0.82
21	0.73	0.73	0.74	0.72	0.75	0.68	0.74
24	0.64	0.64	0.65	0.62	0.65	0.58	0.64
27	0.58	0.57	0.59	0.53	0.59	0.48	0.56
30	0.51	0.50	0.52	0.46	0.53	0.39	0.49
Average Power	0.83	0.83	0.84	0.82	0.84	0.80	0.83

Notes: Table 11 presents power results for CW and different versions of our test when parameters are estimated with a recursive scheme. K is the number of independent realizations of the sequence of θ_t and h is the forecasting horizon. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). The last row reports the average power results across the 30 forecasting horizons. $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\widehat{e_2})$). The total number of Monte Carlo simulations is 2,000 and the sample size is T=900 (R=450 and P=450). We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 10% significance level. Multistep-ahead forecasts are computed using the iterated approach.

Table 12: Power comparisons between CW and WCW tests with nominal size of 10%, considering DGP3 and a small sample.

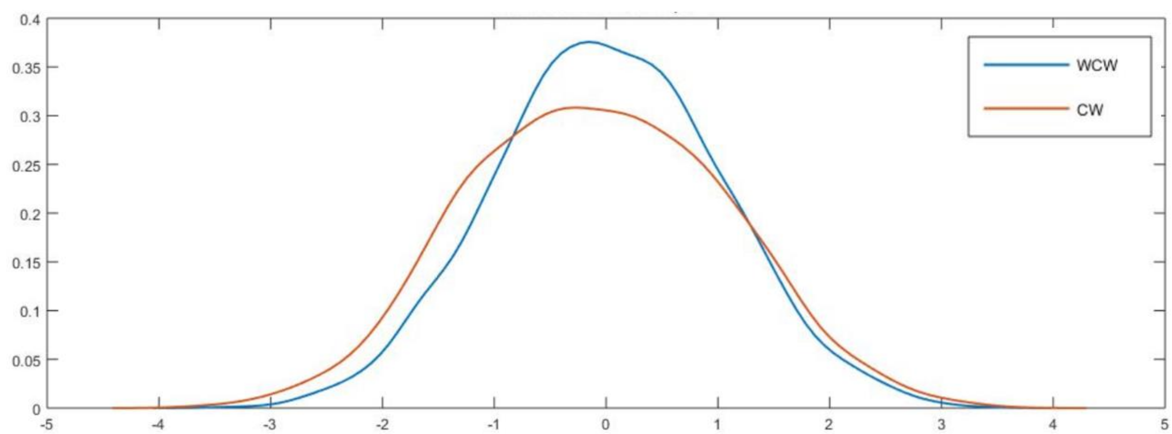
Nominal Size: 0.1		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
h	CW	K=1	K=2	K=1	K=2	K=1	K=2
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	0.98	0.98	0.98	0.98	0.98	0.97	0.98
12	0.73	0.72	0.73	0.71	0.73	0.69	0.73
15	0.65	0.65	0.65	0.63	0.65	0.60	0.64
18	0.60	0.59	0.60	0.57	0.60	0.53	0.58
21	0.55	0.55	0.55	0.53	0.55	0.50	0.53
24	0.52	0.51	0.52	0.49	0.51	0.45	0.50
27	0.49	0.49	0.50	0.46	0.50	0.43	0.48
30	0.48	0.47	0.49	0.45	0.48	0.42	0.46
Average Power	0.70	0.70	0.71	0.69	0.70	0.66	0.70

Notes: Same notes as in Table 11. The only difference is that in Table 10 the sample size is T=180 (R=90 and P=90).

5.3 Simulation results: some comments on asymptotic normality

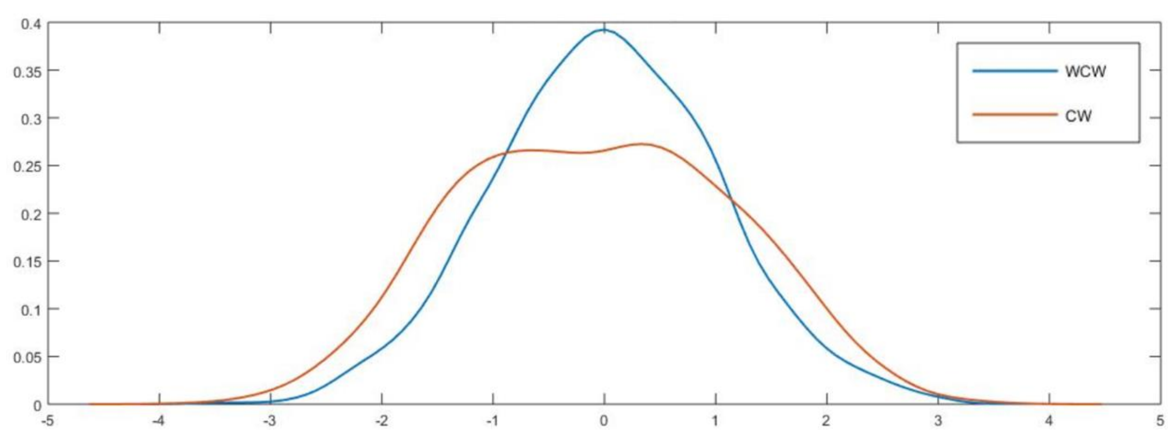
Our simulation exercises show that CW has a pattern of becoming increasingly oversized with the forecasting horizons. At the same time, the WCW tends to have a more "stable" size at long forecasting horizons. These results may, in part, be explained by a substantial departure of normality from CW as h grows. Using DGP2 with h=12, 21 and 27, Figures 1 to 3 support this intuition: while CW shows a strong departure from normality, our WCW seems to behave reasonably well.

Figure 1: Kernel Densities of CW and WCW under the null hypothesis, DGP2, h=12.



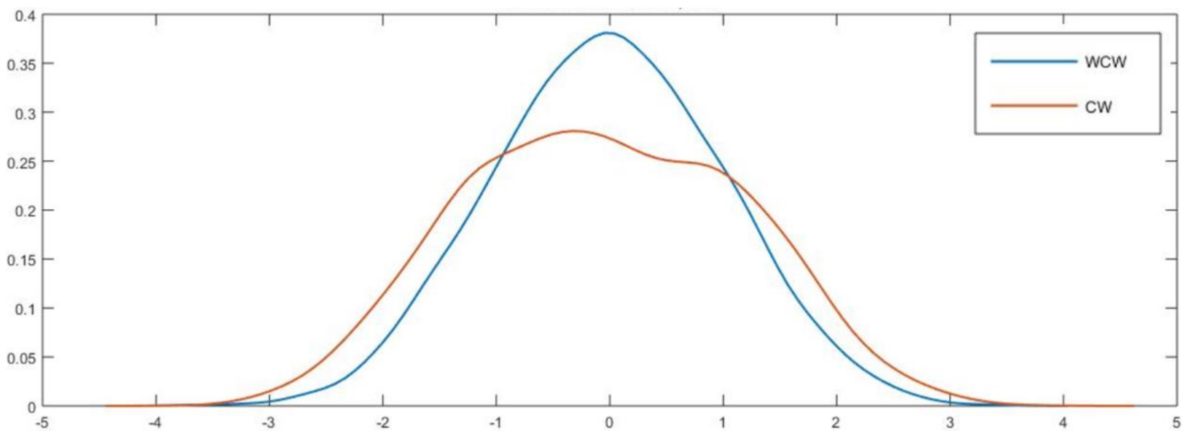
Notes: For this exercise, we consider large samples (P=R=450 and T=900) and 4,000 Monte Carlo Simulations. We evaluate CW and our test computing iterated forecasts. In this case, we use WCW with K=1 and $\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$.

Figure 2: Kernel Densities of CW and WCW under the null hypothesis, DGP2, h=21.



Note: See notes in Figure 1.

Figure 3: Kernel Densities of CW and WCW under the null hypothesis, DGP2, h=27.



Note: See notes in Figure 1.

Table 13 reports the means and the variances of the CW and WCW after 4,000 Monte Carlo simulations. As both statistics are standardized, we should expect means around zero, and variances around one (if asymptotic normality applies). Results in Table 13 are consistent with our previous findings: while the variance of CW is notably high for longer horizons (around 1.5 for h>18), the variance of our test seems to be stable with h, and tends to improve with a higher $\sigma(\theta_t)$. In particular, for the last columns, the average variance of our test ranges from 1.01 to 1.02, and moreover, none of the entries are higher than 1.05 nor lower than 0.98. In sharp contrast, the average variance of CW is 1.32, ranging from 1.07 through 1.51. All in all, these figures are consistent with the fact that WCW es asymptotically normal.

Table 13: Means and Variances of the CW and WCW statistics for DGP2 under the null hypothesis

			$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$				$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$				$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$			
	CW		K=1		K=2		K=1		K=2		K=1		K=2	
h	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
1	-0.16	1.07	-0.06	1.01	-0.10	0.99	-0.01	1.04	-0.04	0.99	0.00	1.05	-0.02	1.00
2	-0.16	1.10	-0.07	1.11	-0.08	1.14	-0.03	1.05	-0.02	1.07	-0.02	1.02	0.01	1.03
3	-0.19	1.08	-0.08	1.12	-0.12	1.18	-0.02	1.04	-0.05	1.09	-0.01	1.00	-0.02	1.04
6	-0.19	1.09	-0.05	1.11	-0.05	1.17	-0.02	1.03	-0.01	1.07	-0.01	1.00	0.01	1.04
9	-0.09	1.19	-0.05	1.07	-0.04	1.10	-0.03	1.03	-0.02	1.05	-0.02	1.02	-0.01	1.03
12	-0.08	1.34	0.00	1.03	-0.05	1.09	0.01	1.00	-0.03	1.04	0.02	1.00	-0.02	1.03
15	-0.06	1.44	-0.04	1.01	-0.02	1.09	-0.03	0.99	0.00	1.04	-0.02	0.99	0.01	1.03
18	-0.06	1.48	-0.04	1.01	-0.03	1.08	-0.03	0.99	-0.01	1.04	-0.02	0.98	-0.01	1.03
21	-0.07	1.51	0.00	1.02	-0.02	1.08	0.01	0.99	0.00	1.03	0.02	0.99	0.01	1.02
24	-0.06	1.51	-0.02	1.07	-0.03	1.08	0.00	1.04	-0.01	1.04	0.00	1.04	0.00	1.03
27	-0.06	1.50	-0.04	1.03	-0.04	1.05	-0.02	1.01	-0.02	1.02	-0.02	1.00	-0.02	1.01
30	-0.06	1.50	-0.01	1.07	-0.03	1.05	0.00	1.05	-0.02	1.00	0.01	1.04	-0.01	0.99
Average	-0.10	1.32	-0.04	1.06	-0.05	1.09	-0.01	1.02	-0.02	1.04	-0.01	1.01	-0.01	1.02

Notes: Table 13 shows the mean and the variance of the CW and WCW statistics after 4,000 Monte Carlo Simulations. For this exercise, we consider large samples (P=R=450 and T=900). We evaluate CW and our test computing iterated forecasts.

6 Empirical Illustration

Our empirical illustration is inspired by the commodity-currencies literature. Relying on the present-value-model for exchange rate determination (Campbell and Shiller (1987) and Engel and West (2005)), Chen, Rogoff and Rossi (2010, 2011), Pincheira and Hardy (2018, 2019a, 2019b) and many others show that the exchange rates of some commodity-exporting countries have the ability to predict the prices of the commodities being exported and other closely related commodities as well.

Based on this evidence, we study the predictive ability of three major commodity-producers economies frequently studied by this literature: Australia, Chile and South Africa. To this end, we consider the following 9 commodities/commodity indices: 1) WTI-Oil, 2) Copper, 3) S&P GSCI: Goldman Sachs Commodity Price Index, 4) Aluminum, 5) Zinc, 6) LMEX: London Metal Exchange Index, 7) Lead, 8) Nickel, and 9) Tin.

The source of our data is Thomson Reuters Datastream, from which we download the daily close price of each asset. Our series are converted to the monthly frequency by sampling from the last day of the month. The time-period of our database goes from September 1999 through June 2019[8].

Our econometric specifications are mainly inspired by Chen, Rogoff and Rossi (2010) and Pincheira and Hardy (2018, 2019a, 2019b). Our null model is

$$\Delta \log(CP_{t+1}) = \mathrm{c}_0 + \rho_0 \Delta \log(CP_t) + \varepsilon_{0,\mathrm{t+1}}$$

While the alternative model is

$$\Delta \log(CP_{t+1}) = \mathrm{c}_1 + \beta \Delta \log(ER_t) + \rho_1 \Delta \log(CP_t) + \varepsilon_{1,\mathrm{t+1}}$$

Where $\Delta \log(C\mathrm{P}_{\mathrm{t+1}})$ denotes the log-difference of a commodity price at time t+1, $\Delta \log(ER_t)$ stands for the log-difference of an exchange rate at time t; c_0, ρ_0 are the regression parameters for the null model and $\mathrm{c}_1, \beta, \rho_1$ are the regression parameters for the alternative model. Finally $\varepsilon_{0,\mathrm{t+1}}$ and $\varepsilon_{1,\mathrm{t+1}}$ are error terms.

One-step-ahead forecasts are constructed in an obvious fashion through both models. Multi-step-ahead forecasts are constructed iteratively for the cumulative returns from t through t+h. To illustrate, let $y_t^f(1)$ be the one-step-ahead forecasts from t to t+1 and $y_{t+1}^f(1)$ the one-step-ahead forecast from t+1 to t+2; then the two-steps-ahead forecast is simply $y_t^f(1) + y_{t+1}^f(1)$.

[8] The starting point of our sample period is determined by the date in which monetary authorities in Chile decided to pursue a pure flotation exchange rate regime.

Under the null hypothesis of equal predictive ability, the exchange rate has no role in predicting commodity prices, i.e $H_0: \beta = 0$. For the construction of our iterated multi-step-ahead forecasts, we assume that $\Delta \log(ER_t)$ follows an AR(1) process. Finally, for our out-of-sample evaluations, we consider P/R=4 and a rolling scheme.

Following eq.(1), we take the adjusted average of K=2 WCW statistics, and consider $\sigma(\theta_t) = 0.04 * \sigma(e_2)$. Additional results using a recursive scheme, other splitting decisions (P and R) and different values of $\sigma(\theta_t)$ and K are available upon request.

Tables 14 and 15 show our results for Chile and Australia respectively. Table A.7 in the Appendix section reports our results for South Africa. Tables 14-15 show interesting results for the LMEX. In particular, the alternative model outperforms the AR(1) for almost every forecasting horizon, using either the Australian Dollar or the Chilean Peso. A similar result is found for aluminum prices when considering $h \geq 3$. These results seems to be consistent with previous findings. For instance, Pincheira and Hardy (2018, 2019a, 2019b), using the ENCNEW test of Clark and McCracken (2001), show that models using exchange rates as predictors generally outperform simple AR(1) processes when predicting some base-metal prices one-step-ahead.

Interestingly, using the Chilean exchange rate, Pincheira and Hardy (2019) report very unstable results at the monthly frequency for nickel and zinc; moreover, they report some exercises in which they could not outperform an AR(1). This is again consistent with our results reported in Table 14.

Results of the CW and our WCW tests are similar. Most of the exercises tend to have the same sign and the statistics have similar "magnitudes." However, there are some important differences worth to be mentioned. In particular, CW tends to reject the null hypothesis more frequently. There are two possible explanations for this result. On the one hand, our simulations reveal that CW has, frequenly, higher power; on the other hand, CW tends to be more oversized than our test at long forecasting horizons, especially for $h \geq 12$. Table 14 can be understood using these two points. Both tests tend to be very similar for short forecast horizons; however, some discrepancies become apparent at longer horizons. Considering $h \geq 12$, CW rejects the null hypothesis at the 10% significance level in 54 out of 81 exercises (67%), while the WCW rejects the null only 42 times (52%). Table 15 has a similar message: CW rejects the null hypothesis at the 5% significance level in 49 out of 81 exercises (60%), while WCW rejects the null only 41 times (51%). The results for Oil (C1) in Table 15 enphasizes this result: CW rejects the null at the 5% significance level for most of the exercises with $h \geq 12$, but our test only rejects at the 10% in most of the exercises. In summary, CW shows a higher rate of rejections at long horizons. The question here is whether this higher rate is due to higher size-adjusted-power, or due to a false discovery rate induced by an empirical size that is higher than the nominal size. While the answer to this question cannot be known for certain, a conservative approach, one that protects the null hypothesis, would suggest to look at these extra CW rejections with caution.

Table 14: Forecasting commodity prices with the Chilean exchange rate. A comparison between CW and WCW in iterated multi-step-ahead forecasts.

Chile CW									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	0.90	0.01	1.24	0.80	-0.33	0.50	0.93	-0.01	0.86
2	0.86	0.68	1.01	1.18	-0.80	1.33*	0.81	-0.19	1.55*
3	0.90	0.06	1.22	1.42*	-0.52	1.67**	1.14	0.97	1.32*
6	0.17	-0.46	0.97	1.42*	-0.77	1.61*	1.51*	1.54*	1.46*
12	0.28	-0.13	0.35	1.68**	-0.95	2.17**	1.27	0.79	1.88**
14	0.65	0.16	0.60	1.71**	-0.98	2.25**	1.20	0.69	1.83**
20	1.24	2.01**	1.33*	1.72**	-1.02	2.26**	1.40*	0.36	1.86**
21	1.29*	1.88**	1.42*	1.71**	-1.02	2.25**	1.39*	0.31	1.77**
22	1.33*	1.74**	1.49*	1.70**	-1.02	2.23**	1.39*	0.27	1.70**
23	1.36*	1.62*	1.54*	1.69**	-1.02	2.22**	1.38*	0.23	1.63*
24	1.39*	1.52*	1.59*	1.68**	-1.02	2.20**	1.38*	0.19	1.58*
25	1.41*	1.45*	1.62*	1.67**	-1.02	2.19**	1.39*	0.16	1.54*
26	1.42*	1.39*	1.65**	1.65**	-1.02	2.17**	1.38*	0.13	1.50*
Chile WCW - K=2 - $\sigma(\theta_t) = 0.04 * \sigma(\widehat{e_2})$									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	0.90	-0.03	1.25	0.80	-0.35	0.51	0.96	0.01	0.86
2	0.86	0.66	1.02	1.19	-0.77	1.38*	0.83	-0.16	1.54*
3	0.90	0.03	1.22	1.41*	-0.49	1.77**	1.14	0.99	1.36*
6	0.19	-0.45	1.00	1.41*	-0.75	1.60*	1.50*	1.54*	1.46*
12	0.34	1.23	0.05	1.68**	-0.94	2.22**	1.26	0.72	1.83**
14	0.36	-0.62	0.45	1.71**	-0.97	2.23**	1.22	0.66	1.86**
20	0.75	1.75**	0.90	1.71**	-1.02	2.28**	1.38*	0.03	2.31**
21	0.92	2.08**	2.36***	1.68**	-1.02	2.23**	1.42*	-0.04	2.14**
22	1.48*	-0.93	1.82**	1.70**	-1.02	2.30**	1.53*	2.52***	1.81**
23	0.69	1.75**	1.48*	1.72**	-1.02	2.26**	1.35*	0.44	1.84**
24	1.94**	1.97**	0.75	1.78**	-1.02	2.19**	1.29*	-0.33	1.11
25	0.23	1.70**	0.79	1.71**	1.02	0.70	1.24	2.21**	0.88
26	1.72**	-1.01	1.38*	1.61*	-1.02	0.90	1.34*	0.51	0.03

Notes: Table 14 shows out-of-sample results using the Chilean exchange rate as a predictor. We report the test by CW and the WCW for P/R=4 using a rolling window scheme. C1 denotes WTI-Oil, C2: Copper, C3: S&P GSCI: Goldman Sachs Commodity Price Index, C4: Aluminum, C5: Zinc, C6: LMEX: London Metal Exchange Index, C7: Lead, C8: Nickel, and C9: Tin. Following eq.(1), we take the adjusted average of K=2 WCW statistics, and we consider $\sigma(\theta_t) = 0.04 * \sigma(\widehat{e_2})$.

Table 15: Forecasting commodity prices with the Australian exchange rate. A comparison between CW and WCW in iterated multi-step-ahead forecasts.

Australia CW									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	-0.51	-0.16	-0.60	-0.65	-0.18	-0.14	1.59*	0.51	0.84
2	0.01	-0.14	-0.68	-0.59	-0.66	0.00	1.05	0.08	1.75**
3	0.41	0.81	0.12	0.11	0.53	1.18	1.49*	1.04	0.99
6	0.72	1.78**	2.13**	1.69**	2.43***	2.25**	1.27	1.74**	1.61*
12	1.55*	1.49*	2.33***	1.74**	1.86**	2.23**	1.28*	2.53***	2.43***
14	1.65**	1.42*	2.278**	1.72**	0.77	2.21**	1.39*	2.39***	2.31**
20	1.68**	1.28*	2.11**	1.68**	-0.81	2.17**	1.46*	1.91**	1.90**
21	1.68**	1.27	2.08**	1.67**	-0.88	2.16**	1.41*	1.85**	1.83**
22	1.67**	1.25	2.05**	1.66**	-0.92	2.15**	1.37*	1.80**	1.76**
23	1.66**	1.24	2.03**	1.65**	-0.95	2.14**	1.33*	1.76**	1.70**
24	1.66**	1.22	2.00**	1.64*	-0.97	2.13**	1.30*	1.72**	1.65**
25	1.65**	1.21	1.98**	1.63*	-0.99	2.13**	1.28*	1.68**	1.60*
26	1.65**	1.20	1.95**	1.62*	-1.00	2.12**	1.25	1.65**	1.55*
Australia WCW - K=2 - $\sigma(\theta_t) = 0.04 * \sigma(e_2)$									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	-0.56	-0.19	-0.61	-0.64	-0.22	-0.14	1.61*	0.54	0.84
2	-0.02	-0.15	-0.68	-0.58	-0.61	0.01	1.06	0.11	1.73**
3	0.42	0.76	0.14	0.07	0.59	1.25	1.46*	1.08	1.04
6	0.75	1.78**	2.13**	1.68**	2.42***	2.25**	1.26	1.74**	1.61*
12	1.55*	1.48*	2.33***	1.73**	2.00**	2.23**	1.28	2.51***	2.42***
14	1.60*	1.43*	2.28**	1.72**	0.74	2.21**	1.38*	2.40***	2.31**
20	1.62*	1.29*	2.11**	1.67**	-0.81	2.17**	1.44*	1.87**	1.95**
21	1.61*	1.27	2.09**	1.63*	-0.90	2.16**	1.45*	1.80**	1.88**
22	1.56*	1.26	2.06**	1.66**	-0.93	2.15**	1.41*	1.96**	1.81**
23	1.76**	1.24	2.04**	1.68**	-0.98	2.14**	1.30*	1.77**	1.69**
24	1.91**	1.22	1.99**	1.74**	-0.98	2.13**	1.27	1.61*	1.61*
25	1.39*	1.22	1.99**	1.74**	1.11	2.13**	1.33*	1.73**	1.55*
26	1.43*	1.22	1.99**	1.49*	-1.01	2.12**	1.16	1.64*	1.52*

Notes: Table 15 shows out-of-sample results using the Australian exchange rate as a predictor. We report the test by CW and the WCW for P/R=4 using a rolling window scheme. C1 denotes WTI-Oil, C2: Copper, C3: S&P GSCI: Goldman Sachs Commodity Price Index, C4: Aluminum, C5: Zinc, C6: LMEX: London Metal Exchange Index, C7: Lead, C8: Nickel, and C9: Tin. Following eq.(1), we take the adjusted average of K=2 WCW statistics, and we consider $\sigma(\theta_t) = 0.04 * \sigma(\widehat{e_2})$.

7 Concluding Remarks

In this paper, we present a new asymptotically normal test for out-of-sample evaluation in the context of nested models. We label this statistic as "Wild Clark and West (WCW)." In essence, we propose a simple modification of the CW (Clark and McCracken (2001) and Clark and West (2006, 2007)) core statistic that ensures asymptotic normality. The key point of our strategy is to introduce a random variable that prevents the CW core statistic from becoming degenerate under the null hypothesis of equal predictive accuracy. Using West (1996) asymptotic theory, we show that "asymptotic irrelevance" applies, hence our test can ignore the effects of parameter uncertainty. As a consequence, our test is extremely simple and easy to implement. This is important since most of the characterizations of the limiting distributions of out-of-sample tests for nested models are non-standard. Additionally, they tend to rely, arguably, on a very specific set of assumptions, that in general, are very difficult to follow by practitioners and scholars. In this context, our test greatly simplifies the discussion in nested models comparisons.

We evaluate the performance of our test (relative to CW), focusing on iterated multi-step-ahead forecasts. Our Monte Carlo simulations suggest that our test is reasonably well-sized in large samples, with mixed results in power compared to CW. Importantly, when CW shows important size distortions at long horizons, our test seems to be less prone to these distortions and therefore it offers a better protection to the null hypothesis.

Finally, based on the commodity currencies literature, we provide an empirical illustration of our test. Following Chen, Rossi and Rogoff (2010,2011) and Pincheira and Hardy (2018, 2019a, 2019b), we evaluate the predictive performance of the exchange rates of three major commodity producers (Australia, Chile and South Africa) when forecasting commodity prices. Consistent with previous literature, we find evidence of predictability for some of our set of commodities. Although both tests tend to be similar, we do find some differences between CW and the WCW. As our test tends to "better protect the null hypothesis," some of these differences may be explained by some size distortions in the CW test at long horizons, but some others are most likely explained by the fact that CW may be, sometimes, more powerful.

Extensions for future research include the evaluation of our test using the direct method to construct multi-step ahead forecasts. Similarly, our approach seems to be flexible enough to be used in the modification of other tests. It should be interesting to explore via simulations its potential when applied to other traditional out-of-sample test of predictive ability in nested environments.

8 References

1. Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3), 817-858.

2. Busetti, F., and Marcucci, J. (2013). Comparing forecast accuracy: a Monte Carlo investigation. *International Journal of Forecasting*, 29(1), 13-27.

3. Campbell, J. Y., and Shiller, R. J. (1987). Cointegration and Tests of Present Value Models. *Journal of Political Economy*, 95(5), 1062–1088.

4. Chen, Y.-C., Rogoff, K. S., and Rossi, B. (2010). Can Exchange Rates Forecast Commodity Prices ? *Quarterly Journal of Economics*, 125(August), 1145–1194.

5. Chen, Y.-C., Rogoff, K. S., and Rossi, B. (2011). Predicting Agri-Commodity Prices: An Asset Pricing Approach, World Uncertainty and the Volatility of Commodity Markets, ed. B. Munier, IOS.

6. Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4), 671-690.

7. Clark, T. E., and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105, 85–110.

8. Clark, T. E., and McCracken, M. W. (2005). The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics*, 124(1), 1–31.

9. Clark, T. E., and West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1–2), 155–186.

10. Clark, T. E., and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.

11. Clark, T., and McCracken, M. (2013a). Advances in forecast evaluation. In Handbook of Economic Forecasting, vol. 2B., Elsevier, Amsterdam, 1107-1201.

12. Clark, T., and McCracken, M. (2013b). Evaluating the accuracy of forecasts from vector autoregressions. In: Fomby. T., Kilian, L., Murphy, A. (Eds.), Vector Autoregressive Modeling – New Developments and Applications: Essays in Honor of Christopher A. Sims, Emerald Group Publishing, Bingley.

13. Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, 12(8), 617–637.

14. Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business and Economic Statistics*, 33(1), 1-1.

15. Diebold, F. X., and Mariano, Roberto, S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statics*, 13(3), 253–263.

16. Engel, C., and West, K. D. (2005). Exchange Rates and Fundamentals. *Journal of Political Economy*, 113(3), 485–517.

17. Giacomini, R., and Rossi, B. (2013). Forecasting in macroeconomics. In Handbook of research methods and applications in empirical macroeconomics, Cheltenham, UK: Edward Elgar Publishing.,Chapter 17, 381-408.

18. Harvey, D. S., Leybourne, S. J., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, 16(2), 254–259.

19. Mankiw, N. G., and Shapiro, M. D. (1986). Do we reject too often?: Small sample properties of tests of rational expectations models. *Economics Letters*, 20(2), 139-145.

20. McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2), 719–752.

21. Nelson, C. R., and Kim, M. J. (1993). Predictable Stock Returns: The Role of Small Sample Bias. *The Journal of Finance*, 48(2), 641–661.

22. Newey, W. K., and West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, 61(4), 631–653.

23. Newey, W. K., and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28(3) ,777-787.

24. Pincheira, P., and Hardy, N. (2019). Forecasting Aluminum Prices with Commodity Currencies Forecasting Aluminum Prices with Commodity Currencies. Available at SSRN 3511564.

25. Pincheira, P., and Hardy, N. (2018). The predictive relationship between exchange rate expectations and base metal prices. Available at SSRN 3263709.

26. Pincheira, P., and Hardy, N. (2019). Forecasting base metal prices with the Chilean exchange rate. *Resources Policy*, 62(February), 256–281.

27. Pincheira, P., and West, K. D. (2016) A comparison of some out-of-sample tests of predictability in iterated multi-step-ahead forecasts. *Research in Economics* 70.2: 304-319.

28. Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54(3), 375-421.

29. West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, 64(5), 1067.

30. West, K. D. (2006). Chapter 3 Forecast Evaluation. Handbook of Economic Forecasting, 1(05), 99–134.

9 Appendix

A.1 Assumption 1 in West (1996) pp.1070:

In some open neighborhood N around β^* , and with probability one:

a) $f_t(\beta)$ is measurable and twice continuously differentiable with respect to β.
b) Let f_{it} be the ith element of f_t. For $i = 1, \dots, l$ there is a constant $D < \infty$ such that for all t, $sup_{\beta \in N} |\frac{\partial^2 f_{it}}{\partial\beta\partial\beta'}| < m_t$ for a measurable m_t for which $\mathbb{E}m_t < D$.

A.2 Assumption 2 on West (1996) pp.1070-1071 and West (2006) pp.112:

Assuming that models are parametric, the estimator of the regression parameters satisfies

$$\widehat{\beta_t} - \beta^* = B(t)H(t),$$

Where $B(t)$ is $k \; x \; q$, $H(t)$ is $q \; x \; 1$ with

a) $B(t) \xrightarrow{a.s} B, B$ a matrix of rank k;
b) $H(t)= t^{-1}\sum_{s=1}^{t} h_s(\beta^*)$ if the estimation method is recursive, $H(t)= R^{-1}\sum_{s=t-R+1}^{t} h_s(\beta^*)$ if it is rolling or $H(t)= R^{-1}\sum_{s=1}^{R} h_s(\beta^*)$ if it is fixed. $h_s(\beta^*)$ is a $q \; x \; 1$ orthogonality condition that satisfies
c) $Eh_s(\beta^*) = 0$.

As it is explained in West (2006): *"Here, h_t can be considered as the score if the estimation method is ML, or the GMM orthogonality condition if GMM is the estimator. The matrix $B(t)$ is the inverse of the Hessian if the estimation method is ML or a linear combination of orthogonality conditions when using GMM, with large sample counterparts B."* West (2006) pp.112.

A.3 Assumption 3 in West (1996) pp.1071:

Let $f_t \equiv f_t(\beta^*)$, $f_{t\beta} \equiv \frac{\partial f_t}{\partial\beta}(\beta^*)$, $F \equiv \mathbb{E}f_{t\beta}$, then

a) For some $d > 1, \sup_t \mathbb{E}| \left|\left[vec(f_{t\beta})', f_t', h_t'\right]'\right|^{4d} < \infty$, where $||\cdot||$ stands for the Euclidean norm.
b) $[vec(f_{t\beta} - F)', (f_t - Ef_t)', h_t']'$ is strong mixing, with mixing coefficients of size $-\frac{3d}{d-1}$
c) $[vec(f_{t\beta})', f_t', h_t']'$ is covariance stationary.
d) $S_{ff} = \sum_{j=-\infty}^{\infty} \Gamma_{ff}(j)$ is p.d. with $\Gamma_{ff}(j) = \mathbb{E}(f_t - Ef_t)(f_{t-j} - \mathbb{E}f_t)'$

A.4 Assumption 4 in West (1996) pp.1071-1072:

$R, P \to \infty$ as $T \to \infty$, and $\lim_{T\to\infty}\left(\frac{P}{R}\right) = \pi, 0 \le \pi \le \infty; \pi = \infty \leftrightarrow \lim_{T\to\infty}\left(\frac{R}{P}\right) = 0$

A.5 Summary of empirical size comparisons between CW and WCW tests with nominal size of 5% for our three DGPs.

Nominal Size: 0.05		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
Average Size	CW	K=1	K=2	K=1	K=2	K=1	K=2
		Large samples (T=900)					
DGP1	0.03	0.04	0.04	0.05	0.05	0.05	0.05
DGP2	0.07	0.05	0.05	0.05	0.05	0.05	0.05
DGP3	0.07	0.06	0.07	0.06	0.06	0.05	0.06
		Small samples (T=90)					
DGP1	0.05	0.05	0.05	0.06	0.06	0.06	0.06
DGP2	0.09	0.07	0.07	0.06	0.07	0.06	0.06
DGP3	0.10	0.08	0.10	0.07	0.09	0.07	0.08

Notes: Table A5 presents a summary of empirical sizes of the CW test and different versions of our test when parameters are estimated with a recursive scheme. Each entry reports the average size across the h=30 excercises. Each row considers a different DGP. The first panel reports our results for large samples (P=R=450, T=900), while the second panel shows our results in small samples (P=R=45, T=90). K is the number of independent realizations of the sequence of θ_t. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000. We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 5% significance level. Multistep-ahead forecasts are computed using the iterated approach.

A.6 Summary of power comparisons between CW and WCW tests with nominal size of 5% for our three DGPs

Nominal Size: 0.05		$\sigma(\theta_t) = 0.01 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.02 * \sigma(\hat{e}_2)$		$\sigma(\theta_t) = 0.04 * \sigma(\hat{e}_2)$	
Average Power	CW	K=1	K=2	K=1	K=2	K=1	K=2
		Large samples (T=900)					
DGP1	0.48	0.47	0.48	0.45	0.47	0.41	0.44
DGP2	0.58	0.48	0.52	0.41	0.45	0.36	0.40
DGP3	0.78	0.78	0.78	0.76	0.79	0.73	0.78
		Small samples (T=90)					
DGP1	0.48	0.47	0.48	0.45	0.47	0.41	0.44
DGP2	0.48	0.43	0.47	0.35	0.41	0.28	0.35
DGP3	0.61	0.61	0.61	0.59	0.61	0.56	0.61

Notes: Table A6 presents a summary of the empirical power of the CW test and different versions of our test when parameters are estimated with a recursive scheme. Each entry reports the average power across the h=30 excercises. Each row considers a different DGP. The first panel reports our results for large samples (P=R=450, T=900), while the second panel shows our results in small samples (P=R=45, T=90). K is the number of independent realizations of the sequence of θ_t. When K>1, our statistic is the adjusted average of the K WCW statistics, as considered in eq(1). $\sigma(\theta_t)$ is the standard deviation of θ_t and it is set as a percentage of the standard deviation of the forecasting errors of model 2 ($\sigma(\hat{e_2})$). The total number of Monte Carlo simulations is 2,000. We evaluate the CW test and our proposed test using one-sided standard normal critical values at the 5% significance level. Multistep-ahead forecasts are computed using the iterated approach.

A.7 Forecasting commodity prices with the South African exchange rate. A comparison between CW and WCW in iterated multi-step-ahead forecasts.

South Africa CW									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	-1.20	-0.11	-0.89	-0.62	-0.92	0.03	-0.74	0.33	0.82
2	-0.45	0.01	-0.93	-0.73	-0.64	0.27	-0.49	-0.13	1.18
3	-0.03	1.08	-0.09	-0.56	0.34	1.31*	0.27	1.44*	0.26
6	1.01	1.52*	2.12**	1.61*	2.09**	2.00**	1.06	2.40***	1.25
12	1.85**	1.41*	2.33***	1.71**	1.96**	2.13**	1.87**	2.50***	1.60*
14	1.85**	1.36*	2.30**	1.69**	1.64*	2.15**	1.98**	2.34***	1.55*
20	1.75**	1.26	2.15**	1.64*	1.13	2.14**	1.65**	2.06**	1.40*
21	1.74**	1.24	2.12**	1.63*	1.10	2.14**	1.60*	2.02**	1.38*
22	1.72**	1.23	2.09**	1.62*	1.08	2.13**	1.56*	1.99**	1.36*
23	1.70**	1.22	2.07**	1.61*	1.06	2.13**	1.53*	1.96**	1.34*
24	1.69**	1.21	2.04**	1.60*	1.05	2.12**	1.50*	1.93**	1.32*
25	1.67**	1.20	2.01**	1.59*	1.04	2.11**	1.47*	1.91**	1.30*
26	1.66**	1.18	1.99**	1.58*	1.04	2.11**	1.44*	1.88**	1.29*
South Africa WCW - K=2 - $\sigma(\theta_t) = 0.04 * \sigma(\hat{e_2})$									
h	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	-1.28	-0.16	-0.90	-0.61	-0.96	0.04	-0.71	0.36	0.82
2	-0.53	0.00	-0.92	-0.72	-0.56	0.29	-0.47	-0.09	1.13
3	0.02	1.02	-0.07	-0.57	0.41	1.39*	0.25	1.46*	0.35
6	1.10	1.53*	2.12**	1.60*	2.10**	2.00**	1.06	2.37***	1.28
12	1.86**	1.40*	2.34***	1.70**	1.96**	2.13**	1.86**	2.49***	1.60*
14	1.80**	1.37*	2.31**	1.69**	1.65**	2.14**	1.98**	2.35***	1.56*
20	1.75**	1.27	2.15**	1.62*	1.13	2.14**	1.64*	2.06**	1.45*
21	1.68**	1.26	2.13**	1.56*	1.10	2.13**	1.61*	2.02**	1.48*
22	1.48*	1.26	2.11**	1.63*	1.08	2.13**	1.57*	1.99**	1.43*
23	1.85**	1.25	2.08**	1.64*	1.06	2.12**	1.52*	1.96**	1.33*
24	1.94**	1.20	2.05**	1.71**	1.05	2.12**	1.49*	1.93**	1.24
25	1.58*	1.31*	1.99**	1.73**	1.04	2.11**	1.47*	1.91**	1.20
26	1.15	1.28	2.06**	1.43*	1.04	2.11**	1.44*	1.88**	1.17

Notes: Table A.7 shows out-of-sample results using the South African exchange rate as a predictor. We report the test by CW and the WCW for P/R=4 using a rolling window scheme. C1 denotes WTI-Oil, C2: Copper, C3: S&P GSCI: Goldman Sachs Commodity Price Index, C4: Aluminum, C5: Zinc, C6: LMEX: London Metal Exchange Index, C7: Lead, C8: Nickel, and C9: Tin. Following eq.(1), we take the adjusted average of K=2 WCW statistics, and we consider $\sigma(\theta_t) = 0.04 * \sigma(\hat{e_2})$.