Multi-benchmark Reality Checks

Ignacio Arbués[*], Ramiro Ledo[†] and Mariano Matilla-García[‡]

March 5, 2015

Abstract

We generalize White's Reality Check (RC) to the case when there is more than one benchmark. This amounts to test the null that the best in a class of models is at least as good as all the models in a second class. This can be of interest, for example, to test Granger-causality without specifying a particular model. We analyze the asymptotic properties of the two variants of the test and propose a Bootstrap to obtain critical values.

Then, we revisit the out-of-sample evidence of predictability of the commodity prices using exchange rates of Chen, Rogoff and Rossi (2010). Univariate models of commodity prices are compared to bivariate models in which exchange rates are included and the null is rejected for some comparisons but not for others. The multi-benchmark RC indicates that there is not evidence of predictability.

Key words: Granger Causality; Out-of-sample forecast; Model Selection; Bootstrap.

JEL classification: F31, C52, C53.

[*] Ministerio de Industria, Energía y Turismo and Instituto Complutense de Análisis Económico. Paseo de la Castellana 160, 28071, Madrid, Spain. Tel: 34 913492219. E-mail: iarbues@minetur.es.

[†] Universidad Complutense de Madrid. Facultad de Ciencias Económicas y Empresariales, Departamento de Economía Aplicada II. Calle General Moscardó 27, apartamento 604, 28020, Madrid, Spain. Tel: 34 629372267, 34 913942635, 34 913942457. E-mail: ramirole@ucm.es

[‡] UNED. Paseo Senda del Rey, 11. Fac. CC. Económicas y Empresariales. 28040, Madrid, Spain. Tel: 34 913987215. E-mail: mmatilla@cee.uned.es. Corresponding author.

1 Introduction

In forecasting literature, we often find comparisons in which a number of models are pitted against a benchmark. We can encounter this when the theory dictates an optimal model. For example, if our premises entail that the returns of a certain asset are unpredictable, we have a geometric random walk model for the price. Thus, the theory can be empirically refuted if the benchmark is actually outperformed by other model. Of course, if we perform comparisons between the benchmark and many alternative models with a sample, there is some probability that the benchmark is beaten just by chance. The RC was proposed by White (2000) as a means to deal with such situations in a systematic way. Specifically, he provided a test for the null hypothesis that the benchmark is optimal. More recently, others (Hansen, 2005; Clark and McCracken, 2012) have developed new tests for this null.

However, there are situations in which there is not a unique benchmark but rather a bunch of them. In the example we present later, we want to determine whether in forecasting a given variable, univariate models can be outperformed by bivariate ones that include lags of a certain covariate. This boils down to a Granger-causality test, but without committing to any specific model in advance. In fact, some applications of RC tests can be described this way (for example Hansen, 2005), but they assume that the matter of choosing one specific univariate benchmark is settled in advance. We could in fact go all the way and dispense with models altogether, using the nonparametric test of Matilla-García, Ruiz Marín and Dore (2014), especially if we are interested in nonlinear predictability.

An RC with several benchmarks can be regarded as a comparison between two sets of models. Then, the null is that the best model of the first set (the one containing the benchmarks) is at least as good as all the models in the second set. Besides the Granger-causality example, we can encounter this situations for example, when assessing whether a new class of models outperforms older, well-known ones. It may be of interest as well to determine if a more complex and costly class of models has actual advantages in terms of forecasting error that compensate for their computational burden (for example, linear vs nonlinear).

In section 2 we present two statistics that can be used to test the generalized null that the best benchmark is as good as the best alternative and discuss their properties, whereas in section 3 we assess the performance of the test by means of Monte Carlo simulations. In section 4 we present an application to the problem of whether the exchange rates help forecasting commodity prices. In particular, we use the data and approach of Chen, Rogoff and Rossi (2010) and show how the multi-benchmark RC can help determine a Granger-causality problem. We conclude in section 5 with some comments.

2 The tests

Subsection 2.1 briefly reviews the most relevant literature on RC tests. Subsection 2.2 introduces the notation and those of the assumptions that are necessary to express precisely the null hypothesis. We will draw heavily on Clark and McCracken (2012 and 2014) so we will adhere as much as possible to their notation and we will refer collectively to both articles as CMC. In 2.3, we will propose the statistics of the test (two variants) and show their asymptotic behavior in 2.4. In 2.5, a procedure to obtain critical values of the test is described.

2.1 Reality checks

White (2000) proposed a test for the null that the benchmark model is as good as any one among a set of alternative models. The test statistic is constructed by taking the following steps: (i) evaluate the out-of-sample forecasting errors with a certain loss function, (ii) calculate the differences between the mean of the benchmark and the means of all the alternative models, and (iii) take the maximum of these differences. Asymptotically (West, 1996), the statistic is approximately distributed as the maximum of correlated normals for large samples. White also showed how to obtain critical values of the test using a version of the stationary bootstrap of Politis and Romano (1994). The Superior Predictive Ability test (Hansen, 2005) is a modification of White's RC in which greater power is obtained by normalizing the mean differences by estimates of their standard deviations.

A problem of these two tests is that they do not work well in a scenario that is of particular interest, namely, when the benchmark is nested in the alternative models (unless the benchmark has no estimated parameters). The asymptotic normality that is a requisite for White's and Hansen's tests can be achieved only by some devices in the nested case. For example, Hansen (2005) uses the approach by Giacomini and White (2006) setting the length of the estimation window constant, so that the variance of the estimated parameters does not vanish asymptotically. This asymptotic theory has some drawbacks, such as the requisite of a small rolling estimation window.

On the other hand, we have a well-developed theory of nested comparisons. The asymptotic theory of these tests for the nested, one-step forecast case was developed in Clark and McCracken (2001) and McCracken (2004). In Clark and McCracken (2005), the theory is adapted to the case of direct multistep forecasts. They also show that the critical values can be obtained by means of a parametric bootstrap.

The theory of nested models comparison and the RC converges in Clark and McCracken (2012), where a test is proposed for the same null hypothesis of the RC, but when the benchmark is nested in all the alternative models. Here, the asymptotic distribution of the test is the maximum of non-standard distributions with nuisance parameters. However, the critical values can be obtained by means of a semi-parametric wild bootstrap.

2.2 Environment and null hypothesis

We enumerate from 1 to n the elements of a set of regressors, that we observe from $t = 1$ to $t = T$. With them, we build different linear models to forecast $y_{t+\tau}$, using information up to time t. Each model can thus be identified with a subset ℓ of $\{1,\ldots,n\}$, so $x_{\ell,t}$ is a $n_\ell \times 1$ vector that comprises the values of the predictors that belong to the set ℓ at time t. The vector that includes all the n predictors is denoted by x_t. We split the time span into an initial estimation window of length R and an out-of-sample span of length P, so that $T = R + P$.

Assumption 1. *For any ℓ, and $t = R, \ldots, R + P - \tau$ we estimate the vector of parameters $\hat{\beta}_{\ell,t}$ by least squares, that is*

$$\hat{\beta}_{\ell,t} = \arg\min_\beta \sum_s^t |y_{s+\tau} - x'_{\ell,s}\beta|^2,$$

where the s ranges from 1 to t in the 'recursive' scheme and from $t - R + 1$ to t in the 'rolling' scheme.

With the estimated parameters, we build forecasts $\hat{y}_{i,t+\tau} = x'_{i,t}\hat{\beta}_{i,t}$ for $t = R, \ldots, T - \tau$. The forecast errors are $\hat{u}_{i,t+\tau} = y_{t+\tau} - \hat{y}_{i,t+\tau}$. Among the 2^n possible models, we will restrict the analysis to the elements of two particular classes, I and J, that is, the class of the benchmarks and the class of the alternatives. Generally, i and j will denote models in I and J respectively, whereas ℓ indicates models that may belong either to I or to J.

The performance of each model is measured in terms of Mean Squared Forecasting Error (MSFE). Let $\sigma^2_\tau(\ell)$ be the population MSFE of model ℓ, that is, $\mathbb{E}(u_{\ell,t+\tau})^2$, where $u_{\ell,t+\tau} = y_{t+\tau} - x'_{\ell,t}\beta^*_\ell$ and β^*_ℓ is the population parameter vector of model ℓ. The population forecasts errors of the full model are $u_t = y_{t+\tau} - x'_t\beta^*$ and the MSFE is $\mathbb{E}u^2_{t+\tau} = \sigma^2_\tau$.

Now, we can state the null hypothesis as

$$H_0 : \min_{i \in I} \sigma^2_\tau(i) \leq \min_{j \in J} \sigma^2_\tau(j),$$

that is, no model in J beats the best model in I.

2.3 Test statistics

The condition $\min_i \sigma^2_\tau(i) \leq \min_j \sigma^2_\tau(j)$ is equivalent to either $\min_i \max_j(\sigma^2_\tau(i) - \sigma^2_\tau(j)) \leq 0$ or $\max_j \min_i(\sigma^2_\tau(i) - \sigma^2_\tau(j)) \leq 0$. This way, we express H_0 in terms of pairwise comparisons. Since we have tests for these comparisons, we may combine them into a statistic for the whole multiple comparison.

In order to specify the statistics we need to introduce some further notation. Let $\hat{d}_{ij,t} = \hat{u}^2_{i,t+\tau} - \hat{u}^2_{j,t+\tau}$, $\bar{d}_{ij} = (P-\tau+1)^{-1}\sum_{t=R}^{R+P-\tau}\hat{d}_{ij,t}$ and $\hat{\gamma}_{d_{ij}}(l) = (P-\tau+1)^{-1}\sum_{t=R+l}^{R+P-\tau}(\hat{d}_{ij,t}-\bar{d}_{ij})(\hat{d}_{ij,t-l}-\bar{d}_{ij})$. We can use the covariances $\hat{\gamma}_{d_{ij}}(l)$ to estimate the long-run covariance of $\bar{d}_{ij}$ as $(P-\tau+1)^{-1/2}\hat{S}_{d_{ij},d_{ij}}$, where $\hat{S}_{d_{ij},d_{ij}} = \sum_{l=-\bar{l}}^{\bar{l}} K(l/L)\hat{\gamma}_{d_{ij}}(l)$, $K(\cdot)$ is a certain kernel, L is a truncation parameter and $\hat{\gamma}_{d_{ij}}(-l) = \hat{\gamma}_{d_{ij}}(l)$. We also write $\hat{\sigma}^2_\tau(\ell) = (P-\tau+1)^{-1}\sum_{t=R}^{R+P-\tau}\hat{u}^2_{\ell,t+\tau}$.

We build our test from pairwise comparison statistics. However, note that when comparing $i \in I$ and $j \in J$, we will have both cases where the models nested and nonnested. Then, the pairwise comparison statistics will be either $\Delta_{ij} = \text{MSE-t}_{ij} = (P-\tau+1)^{1/2}\bar{d}_{ij}/\sqrt{\hat{S}_{d_{ij},d_{ij}}}$ for nonnested models or $\Delta_{ij} = \text{ENC-F}_{ij} = (P-\tau+1)\bar{d}_{ij}/\hat{\sigma}^2_\tau(j)$ when i is nested in j. We use MSE-t when the models are nonnested because is symmetric and ENC-F when the models are nested because it is apparently the most powerful according to the simulations reported in Clark and McCracken (2012). Finally, the test statistics are

$$\Delta\text{-mM} = \min_{i\in I}\max_{j\in J}\Delta_{ij} \tag{1}$$

$$\Delta\text{-Mm} = \max_{j\in J}\min_{i\in I}\Delta_{ij}. \tag{2}$$

When there is only one element in I, both statistics collapse to the $\max_j$ ENC-F test of Clark and McCracken (2012).

2.4 Asymptotic properties

To determine the asymptotic properties of the test requires to make some assumptions. A possibility that requires relatively mild assumptions is to invoke the conditional approach of Giacomini and White (2006), as Hansen (2005) does for his RC. Under this framework, the asymptotic distribution of the pairwise statistics Δ_{ij} is normal. However, one has to pay the price of assuming a fixed-length estimation window R. There is an argument that suggests that in practice R should be quite small. Suppose you applied the test in the past, when the series were short. Then, necessarily R had to be small. Later, with more observations, you should not take advantage of the greater length of the series, because R is assumed fixed.

This inconvenience notwithstanding, we intend to explore this possibility in other article. Here we will adapt instead the asymptotic theory and the bootstrap of Clark and McCracken (2012, 2014). Although this requires to impose some restrictions on which kind of models can be compared and other assumptions, there are still many applications that satisfy the requirements.

We will make two kinds of assumptions. First, we will impose some restrictions on the set-theoretical relationship between I and J. In the RC by CMC, the restriction is very simple: the benchmark is nested in every alternative. Here, the assumption has to be more involved. The second set of assumptions is directly taken from CMC are relate to the statistic properties of the processes.

We want to preserve one idea of the single-benchmark RC, namely, that the benchmark is a simpler model that is compared to more complex ones. However, the assumption that all $i \in I$ are nested in all $j \in J$ would be too strong. For example, when testing causality of a variable z_t on y_t, I contains models with only lags of y_t, whereas the models in J use as well lags of z_t. In this case, in order that i is nested in j for all $i \in I, j \in J$, it would be necessary to force that all bivariate models include as many lags of y_t as the largest univariate model, which is not necessarily what we intend. However, it is necessary to introduce a restriction, so that we can simplify the asymptotic distribution.

In plain words, we will require that there is a model k_0 that nests all benchmarks, but not any alternative; that all benchmarks are nested in at least one alternative (so the test is one-sided); that there alternatives enough to check the predictive ability of all regressors in k_0^c; and conversely that the predictive gains of the alternatives are due only to the regressors in k_0^c. We state this in mathematical terms.

Assumption 2. *There is a* $k_0 \subsetneq \{1, \ldots, n\}$ *that contains every* $i \in I$ *and such that (a)* $\forall i \in I, a \in k_0^c, \exists j \in J$ *such that* $i \cup \{a\} \subseteq j$ *and (b)* $\forall j \in J, \exists i \in I$ *such* *that $j \cap k_0 \subset i$.*

In the case of Granger causality, k_0 is the set of the lags of the predictand and k_0^c has lags of another variable. Assumption 2 is satisfied, for example, when I is any set of autoregressive models and J includes models that have the same lags of y_t than some of I, plus a certain number of lags of the regressor z_t. This restriction gives space, for example, for: (i) univariate models vs multivariate (or in general, $p-$variate vs $q-$variate, with $p < q$); (ii) linear vs nonlinear (not any nonlinear model, but polynomial models can be accommodated); (iii) forecasting with aggregate data vs aggregating forecasts.

We define I_0 and J_0 as the sets of models that minimize $\sigma_\tau^2(\ell)$ in I and J respectively. That is, I_0 comprises the "good" models of I and J_0 the "good" models of J. Under the null, all of the models in $I_0 \cup J_0$ have the same population MSFE, whereas the remaining ones –the "bad" ones– that belong to $(I \setminus I_0) \cup (J \setminus J_0)$ have greater MSFE.

We need some additional notation for the asymptotic analysis: J_ℓ is the $n_\ell \times n$ selection matrix such that $x_{\ell,t} = J_\ell x_t$; $B = (\mathbb{E}x_t x_t')^{-1}$, $B_\ell = (\mathbb{E}x_{\ell,t}x_{\ell,t}')^{-1}$; $B(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_s x_s')^{-1}$, $B_\ell(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{\ell,s}x_{\ell,s}')^{-1}$.

Let $h_t = x_t u_t$, $H(t) = t^{-1}\sum_{s=1}^{t-\tau} h_t$. Let $\tilde{A}_{ij}$ be a $n \times n$ matrix with rank $n_i + n_j - 2n_{i\cap j}$ such that $\tilde{A}_{ij}'\tilde{A}_{ij} = B^{-1/2}(-J_i B_i J_i' + J_j B_j J_j')B^{-1/2}$ and $\tilde{h}_{ij,t} = \sigma_\tau^{-1}\tilde{A}_{ij}B^{1/2}h_t$. Let S_{hh} be equal to $\sum_{k=-\tau+1}^{\tau-1}\Gamma_{hh}(k)$, and $\Gamma_{hh}(k)$ is the autocovariance function of h_t. Finally, $S_{ij} = \sigma_\tau^{-2}S_{hh}^{1/2}B^{1/2}\tilde{A}_{ij}'\tilde{A}_{ij}B^{1/2}S_{hh}^{1/2}$

Further technical assumptions required are the following.

Assumption 3. *(a) $U_{t+\tau} = [h_{t+\tau}', \mathrm{vec}(x_t x_t' - \mathbb{E}x_t x_t')']'$ is covariance stationary. (b) $\mathbb{E}U_{t+\tau} = 0$. (c) For all $l > \tau - 1, \mathbb{E}h_{t+\tau}h_{t+\tau-l}' = 0$. (d) $\mathbb{E}x_t x_t' < \infty$ and is positive definite. (e) for some $r > 8$, $U_{t+\tau}$ is uniformly bounded in L^r. (f) For some $r > d > 2, U_{t+\tau}$ is strong mixing with coefficients of size $-rd/(r-d)$. (g) $\lim_R R^{-1}\mathbb{E}(\sum_{s=1}^{R-\tau}U_{s+\tau})(\sum_{s=1}^{R-\tau}U_{s+\tau})' = \Omega < \infty$ is positive definite.*

Assumption 4. *(a) let $K(x)$ be a continuous kernel such that for all real scalars x, $|K(x)| \leq 1, K(x) = K(-x)$ and $K(0) = 1$. (b) For some bandwidth L and* *constant $i \in (0, 0.5), L = O(P^i)$. (c) The number of covariance terms $\bar{l}$ used to estimate the long-run covariances $S_{d_{ij},d_{ij}}$ satisfies $\tau - 1 \leq \bar{l} < \infty$.*

Assumption 5. $\lim_{P,R} P/R = \lambda_P \in (0, \infty)$.

These assumptions, as many in the literature of this area, are variations around the initial setting of West (1996).

The first difficulty that arises when analyzing the asymptotic behavior of these tests is that they involve one-to-one comparisons of very different kind:

(A) $\sigma^2_\tau(i) < \sigma^2_\tau(j)$.

(B) $\sigma^2_\tau(i) > \sigma^2_\tau(j)$.

(C) $\sigma^2_\tau(i) = \sigma^2_\tau(j)$ and,

- (C1) i is nested in j (the reverse case is possible only in case that $I \cap J \neq \emptyset$, which can be discarded without loss of generality).
- (C2) i and j are overlapping in the sense of Vuong (1989), that is, both contain the true model plus terms that vanish for the population value of the parameters.
- (C3) i and j are nonnested.

When we compare one of the ”good” models to one of the ”bad” ones, the corresponding MSE-t$_{ij}$ statistic diverges to either $-\infty$ or $+\infty$. A consequence of this is that in (1) and (2), only the Δ_{ij} with $i \in I_0$ and $j \in J_0$ are asymptotically relevant.

Proposition 1. *Under H_0, Δ-mM $= \min_{i \in I_0} \max_{j \in J_0} \Delta_{ij} + o_p(1)$ and Δ-Mm $= \max_{j \in J_0} \min_{i \in I_0} \Delta_{ij} + o_p(1)$.*

In other words, only the good models matter asymptotically and thus, under the null, cases A and B can be disregarded. Moreover, the same happens with case C3.

We need the following auxiliary lemma, whose proof is left to the reader.

Lemma 1. *For $z \in \mathbb{R}$, let $[z]_+$ be equal to* $\max\{0, z\}$. *If $[-x_t]_+ = O_p(1)$ and $y_t \xrightarrow{p} -\infty$, then* $\max\{x_t, y_t\} - x_t = o_p(1)$.

Proposition 2. *If i and j are in case C3, then $i \notin I_0, j \notin J_0$.*

These propositions entail that we can focus on the asymptotic behavior of Δ_{ij} with $(i, j) \in I_0 \times J_0$, and that this is given by the following results[1]:

(a) When i and j are nested, the asymptotic distribution of $\Delta_{ij} =$ ENC-F is given by theorem 3.2 in Clark and McCracken (2012).

(b) When i and j are overlapping, the asymptotic distribution of $\Delta_{ij} =$ MSE-t is given by theorem 2.1 in Clark and McCracken (2014).

We just need to adapt these results to our framework.

Proposition 3. *The asymptotic distributions of the tests are given by*

$$\Delta\text{-mM} \xrightarrow{d} \min_{i \in I_0} \max_{j \in J_0} g_{ij}, \tag{3}$$

$$\Delta\text{-Mm} \xrightarrow{d} \max_{j \in J_0} \min_{i \in I_0} g_{ij}, \tag{4}$$

where $g_{ij} \sim (\Gamma_{1,ij} - 0.5\Gamma_{2,ij})/\Gamma_{3,ij}^{1/2}$,

$$\Gamma_{1,ij} = \int_\lambda^1 \omega^{-1} W(\omega)' S_{ij} dW(\omega)$$
$$\Gamma_{2,ij} = \int_\lambda^1 \omega^{-2} W(\omega)' S_{ij} W(\omega) d\omega$$
$$\Gamma_{3,ij} = \int_\lambda^1 \omega^{-2} W(\omega)' S_{ij}^2 W(\omega) d\omega,$$

$W(\omega)$ is a $n \times 1$ standard Brownian motion and the S_{ij} are matrices defined below.

When the null does not hold, both statistics diverge to $+\infty$.

Proposition 4. *If the null does not hold,* $\text{plim}\Delta\text{-Mm} = \text{plim}\Delta\text{-mM} = +\infty$.

[1] The theoretical results are developed only for the recursive window, as in CMC. However, there is substantial evidence that the bootstrap works with a rolling window as well. This evidence comes both from our simulations and from CMC, which report a similar behaviour with the two windows.

2.5 The bootstrap

The asymptotic distributions of proposition 3 are impractical to obtain critical values or p-values. Therefore, we will use a bootstrap that is adapted from CMC. They generate artificial samples according to a model,

$$y^*_{t+\tau} = \hat{\beta}_0 x_{0,t} + \hat{v}^*_t,$$

where the vector $x_{0,t}$ contains either the predictors that are common to the two models compared (in the overlapping models test) or the predictors of the benchmark (in the RC). The $\hat{v}^*_t$ terms are simulated by a wild bootstrap designed to retain some features of the true prediction errors such as the heteroskedasticity and when $\tau > 1$ also the autocorrelation. We will take advantage of the fact that under the null, all coefficients of β^* outside k_0 are zero, so we will generate our artificial samples according to $y^*_{t+\tau} = \hat{\beta}_{k_0} x_{k_0,t} + \hat{v}^*_t$.

To be more specific, the following steps are taken.

1 We fit the model with all n regressors and obtain the forecast errors $\hat{v}_t$, with $t = 1, \ldots, R + P - \tau$.

2 We estimate a MA$(\tau - 1)$ model $v_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_{\tau-1}\varepsilon_{t-\tau+1}$, obtaining the residuals $\hat{\varepsilon}_t$.

3 We simulate i.i.d variables η_t and calculate $\hat{v}^*_t = \eta_t \hat{\varepsilon}_t + \hat{\theta}_1 \eta_{t-1} \hat{\varepsilon}_{t-1} + \ldots + \hat{\theta}_{\tau-1} \eta_{t-\tau+1} \hat{\varepsilon}_{t-\tau+1}$.

4 We estimate the parameter β_{k_0} of the model with $x_{k_0,t}$ and build the bootstrapped data $y^*_{t+\tau} = \hat{\beta}_{k_0} x_{k_0,t} + \hat{v}^*_t$.

5 With the sample $y^*_1, \ldots, y^*_{R+P-\tau}$, calculate the statistics Δ-mM* and Δ-Mm*.

In step 3 we depart from CMC in one respect. We replace the normal distribution used to generate η_t by one among two discrete distributions that take either the values $(-(\sqrt{5}-1)/2, (\sqrt{5}+1)/2)$ with probabilities $p = (\sqrt{5}+1)/(2\sqrt{5})$

and $1-p$ respectively or $(-1,1)$ with probabilities $(0.5,0.5)$. The first distribution satisfies $\mathbb{E}\eta^3 = 1$ and the second $\mathbb{E}\eta^3 = 0, \mathbb{E}\eta^4 = 1$, so they preserve the third and fourth-order moments of $\hat{\varepsilon}_t$ respectively. These distributions are discussed, for example, in Davidson and Flachaire (2008). In our Monte Carlo experiment, we have observed that for finite samples, the empirical sizes obtained with the second distribution were more approximate to their theoretical values. This is consistent with the fact that we simulate the innovations of the prediction error with a symmetric distribution, so the third-order moment preserved even when $\mathbb{E}\eta^3 = 0$. If one has reasons to believe that the innovations are skewed, the distribution with the unitary third-order moment should be chosen instead.

In order to prove the validity of the bootstrap under the null hypothesis, we have to prove that with the bootstrap-induced probability distribution, P^*, the matrix $\{\Delta_{ij}\}_{(i,j)\in I_0\times J_0}$ and its bootstrapped counterpart $\{\Delta^*_{ij}\}_{(i,j)\in I_0\times J_0}$ have the same asymptotic distribution, conditional to the sample. We indicate this convergence by $\overset{d^*}{\to}$. For this, we have to consider the consequences of generating the artificial sample y^*_t with the full set of regressors k_0 instead of either i for the nested case, or $i\cap j$ for the overlapping case. In proposition 5 below we show that this effect is asymptotically negligible. Here we only outline the argument: under the null, if $(i,j)\in I_0\times J_0$, then the regressors with nonzero coefficients are all included in i and j. This means that the excess parameters in $\beta^*_{k_0}$ compared to β^*_i in the nested case or to $\beta^*_{i\cap j}$ in the overlapping case are zero. Since the bootstrap distributions obtained with the regressors i or $i\cap j$ are correct for each case, then so are those obtained with k_0.

In order to prove the validity of the bootstrap we need strengthened versions of assumptions 2 and 3.

Assumption 2′. *There is a $k_0 \subsetneq \{1,\dots,n\}$ that contains every $i\in I$ and such that (a) $\forall i\in I, a\in k_0^c, \exists j\in J$ such that $i\cup\{a\}\subseteq j$, (b) $\forall j\in J, \exists i\in I$ such that $j\cap k_0\subset i$ and (c) $k_0\in I$.*

The new condition (c) ensures that the models in I_0 contain all the relevant regressors. This makes the proof of the validity of the bootstrap much simpler, but it is likely not necessary. We have simulated a scenario in which $k_0 \notin I$ and the tests work properly, which reinforces this conjecture.

Assumption 3$'$. *(a)* $U_{t+\tau} = [h'_{t+\tau}, \mathrm{vec}(x_t x'_t - \mathbb{E}x_t x'_t)']'$ *is covariance stationary. (b)* $\mathbb{E}[\varepsilon_{t+\tau} | \varepsilon_{t+\tau-j}, x_{s-j} : j \geq 0] = 0$*. (c) Let* $\gamma = (\beta', \theta_1, \ldots, \theta_{\tau-1})'$, $\hat{\gamma} = (\hat{\beta}', \hat{\theta}_1, \ldots, \hat{\theta}_{\tau-1})'$ *and define the function* $\hat{\varepsilon}_{s+\tau} = \hat{\varepsilon}_{s+\tau}(\hat{\gamma})$ *such that* $\hat{\varepsilon}_{s+\tau}(\gamma) = \varepsilon_{s+\tau}$*. In an open neighborhood* N *around* γ*, there exists* $r > 8$ *such that* $\sup_{l \leq s \leq R} \| \sup_{\gamma \in N} \hat{\varepsilon}_{s+\tau}(\gamma), \nabla \hat{\varepsilon}'_{s+\tau}(\gamma), x_s)' \|_r \leq c$*. (d)* $\mathbb{E}x_t x'_t < \infty$ *and is positive definite. (e) For some* $r > d > 2, U_{t+\tau}$ *is strong mixing with coefficients of size* $-rd/(r-d)$*. (f)* $\lim_R R^{-1} \mathbb{E}(\sum_{s=1}^{R-\tau} U_{s+\tau})(\sum_{s=1}^{R-\tau} U_{s+\tau})' = \Omega < \infty$ *is positive definite.*

Assumption 3’ is intended to ensure that the process of generating the bootstrap errors $\hat{v}_t^*$ reproduce the behavior of the true errors u_t to the extent required for the asymptotic results.

Proposition 5. *Under* H_0*,*

(i) $\Delta\text{-mM}^* = \min_{i \in I_0} \max_{j \in J_0} \Delta^*_{ij} + o_p(1)$.

(ii) $\Delta\text{-Mm}^* = \max_{j \in J_0} \min_{i \in I_0} \Delta^*_{ij} + o_p(1)$.

(iii) $\{\Delta^*_{ij}\}_{(i,j) \in I_0 \times J_0} \overset{d^*}{\to} \{g^*_{ij}\}_{(i,j) \in I_0 \times J_0}$*, where* $\{g^*_{ij}\}_{(i,j) \in I_0 \times J_0}$ *is distributed as* $\{g_{ij}\}_{(i,j) \in I_0 \times J_0}$*.*

When H_0 does not hold, $\Delta_{ij} \overset{d^*}{\to} g^*_{ij}$, where g^*_{ij} is distributed as the g_{ij} corresponding to $\tilde{y}_{t+\tau} = x'_{k_0,t} \beta_{k_0} + u_t$. With $\tilde{y}_t$, the sets I and J satisfy the null hypothesis because of assumption 2(b). Hence, when the null does not hold,

$$P[\Delta\text{-Mm} > \xi^*_{Mm,\alpha}], P[\Delta\text{-mM} > \xi^*_{mM,\alpha}] \to 1,$$

where $\xi^*_{Mm,\alpha}$ and $\xi^*_{mM,\alpha}$ are the Bootstrap critical values.

3 Monte Carlo

The purpose of our simulations is twofold. In the first place, we want to check that the empirical size of the test is near enough to the theoretical one. On the other hand, we want to compare its power with some other procedures. Instead of our test, one could pick one model from each set and apply a pairwise comparison. For this purpose, we have used both the AIC and BIC criteria. A rather empirical method we included in our experiment is as follows: pick the largest model in I and apply an RC of all models in J against this unique benchmark.

Summarizing, we have the following tests:

(1a, 1b) Select the benchmark with AIC (1a) or BIC (1b).

(2) Select the largest benchmark. For example if we consider the univariate models RW, RWD, AR(1), we use AR(1) as benchmark.

(3a, 3b) Use one of two tests based on

$$\min_i \max_j \Delta_{ij} \qquad \text{and} \qquad \max_j \min_i \Delta_{ij},$$

where Δ_{ij} is for each univariate model i and bivariate model j, either Clark and McCracken's (2001) ENC-F or MSE-T depending on whether i is nested in j or not.

We have run simulations in which bivariate data are generated according to a VAR(3) model. Depending on a parameter b, we make one component to Granger-Cause the other one. We used both the rolling and recursive schemes. The model has the form

$$x_t = 0.6x_{t-1} - 0.3x_{t-2} + 0.2x_{t-3} + \varepsilon_{1t} \tag{5}$$

$$y_t = 0.4y_{t-1} + 0.2y_{t-2} + 0.1y_{t-3} + bx_{t-1} + \varepsilon_{2t}. \tag{6}$$

When $b = 0$, the null holds and the p-values should be uniformly distributed in $[0, 1]$. We test the lack of Granger-Causality using as benchmarks univariate AR(p) and the alternatives are VAR(p) with $p = 1, \ldots, 5$.

In figure I, we see that the empirical cumulative distribution function of the AIC and BIC methods are much above the diagonal, which means that actually they tend to reject too often. On the other hand, the maximin and minimax tests are very close to the diagonal.

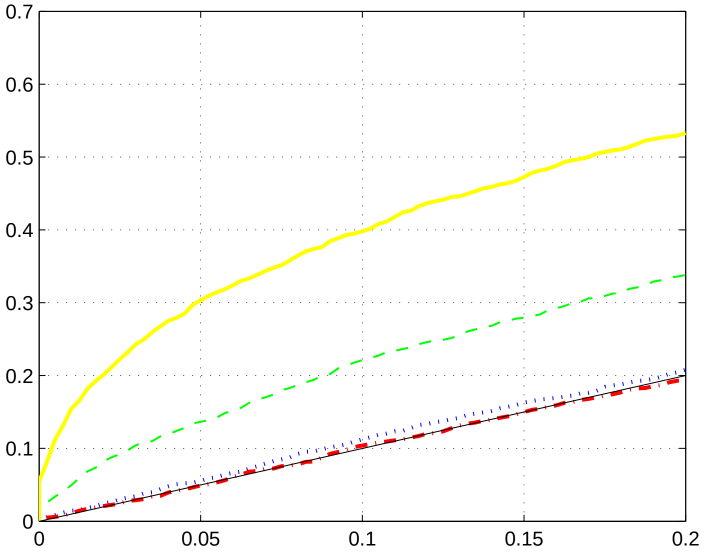

Figure I: Case $b = 0$, null hypothesis, T=160. We represent the empirical distribution functions of the p-values from top to bottom as: method 1a=thick continuous curve; method 1b=dashed curve; minimax=dotted curve; maximin=dash-dot curve. For reference, we draw the diagonal as a thin continuous line. Rolling scheme.

We could expect method 2 to have excessive size sometimes, in situations when the largest benchmark has many parameters and most of them are null in population terms. Then, the estimation variability would increase the forecasting error and make the alternatives look better. However, in our experimentation, we have found this effect to be small. We have designed experiments to discard this method (and thus, to make more appealing our new tests Δ-Mm and Δ-mM), but to no avail.

One of the experiments we have tried is based on the data generating process

$$y_{t+1} = 0.3y_t + b'x_{1t} + \varepsilon_{t+1},$$

where $b = 0, 0.2$, $X_t = (x_{1t}, \ldots, x_{8t})'$ is a multivariate normal vector with a symmetric Toeplitz covariance matrix with first row $(1, 0.9, \ldots, 0.3)$ and ε_{t+1} is an independent standard normal. We consider all the $2^8 = 512$ combinations of regressors. We want to test whether any x_{it} with $i = 1, \ldots 4$ has predictive power. Hence, our benchmarks are the 16 models that do not contain any of the first four regressors and the alternatives are the remaining 240 that do.

When $b = 0$, that is, when the null holds, method 2 has a rejection probability only a little greater than our tests and essentially as the theoretic size. When $b_1 = 0.2$ the power is a little greater as well.

To save space, only a small selection of results were included. We have found generally, quite similar size and power between methods 2, 3a and 3b. We invite the reader to experiment with our MATLAB programs.

4 Exchange rates and commodity prices

There are many models in the literature that relate exchange rates to economic fundamentals and much work has been invested trying to find empirical evidence of their validity. Unfortunately, this has proved a hard task. For example, Meese and Rogoff (1983a, b) conclude that random walk forecasts are as good as any. Later studies, such as Cheung, Chinn and García Pascual (2005), with a new generation of models do not change substantially the picture.

A possible explanation of the failure to measure relationships between exchange rates and other variables is advanced in Engel and West (2004). They assume a present-value model and show that when the dynamics of at least one fundamental has a unit root, then the exchange rates behave approximately as a random walk as the discount factor approaches unity. In view of this, it may not be surprising that in empirical analysis the random walk performs as well as

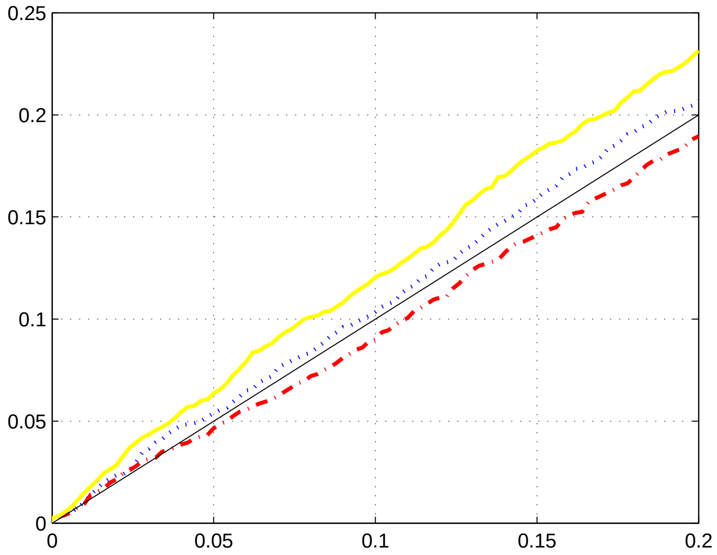

Figure II: Case $b = 0$, null hypothesis, T=80. We represent the empirical distribution functions of the p-values with: method 2=thick continuous curve; minimax=dotted curve; maximin=dash-dot curve. For reference, we draw the diagonal as a thin continuous line. Recursive scheme.

more complex models. They also suggest that present-value models also imply that exchange rates could be useful to predict the fundamentals. The rationale for this is that if current exchange rates depend on discounted expectations of fundamentals, then information related to the future evolution of the latter is already incorporated in exchange rates. It is also possible that this information is not at hand to be used explicitly in econometric models to forecast the fundamentals.

This idea is pursued in Chen, Rogoff and Rossi (2010, hereafter CRR). In particular, they focus on commodity prices. With other fundamentals it is difficult to discard the possibility that exchange rates affect them through other channels than the expectations. For example, monetary policy responses to variations in exchange rates can affect other fundamentals. By avoiding this, we ensure that Granger-causality from exchange rates can be interpreted as evidence for the present-value model.

In CRR, the predictive ability of the exchange rates for the commodity prices is tested through simple linear models for five countries. As in Meese and Rogoff (1983a), the models are evaluated by means of out-of-sample forecast error. Given that the model comparisons are nested, the encompassing tests by Clark and McCracken (2001) are a natural choice. In particular, the ENC-F test appears to reject in most cases the null that the univariate models for commodity prices are as good as the bivariate models including exchange rates. However, we show that there is a bug in the implementation of the test and the results after the correction are not at all clear. All this is discussed in section 4.1.

There is a difficulty in the interpretation of the tests, because several univariate and bivariate models are used. This entails that we have a number of pairwise comparisons. If the null is rejected in some of them, but not all, we need a procedure to arrive at a global decision about whether the bivariate models are collectively an improvement with respect to the univariate models or not. We discuss in section 4.3 some methods to do this. By Monte Carlo simulations, we can discard some of them and reduce the possibilities to only two. Both of them fail to find evidence that the exchange rates are actually useful to predict commodity prices.

4.1 The problem

Let us consider a present-value model such as

$$s_t = \sum_{j=0}^{\infty} \psi^j \mathbb{E}[a' f_{t+j} | I_t],$$

where s_t is the price of an asset, ψ a discount factor, f_t is a vector of fundamentals, I_t is the information set at time t, and a is a vector of coefficients. The objective of the exercise is to check whether such a model could be valid for exchange rates. The route to this is to see if s_t Granger-causes f_t. In particular, we focus on the commodity prices to avoid being entangled with reverse causality.

We consider quarterly series of commodity prices (cp_t), exchange rates relative to dollar (erd_t) and British pound (erp_t) and nominal effective exchange rates (ern_t) from Australia, New Zealand, Canada, Chile and South Africa. They are taken from the data set by CRR retrieved from the personal website of Barbara Rossi.

We want to determine whether the knowledge of erd_t, erp_t or ern_t allows making better predictions of $\mathrm{cp}_{t+\tau}$ than with the latter alone. We focus first in $\tau = 1$, but also include some results later for $\tau = 4$. This can be regarded as a comparison between univariate and bivariate models. For simplicity, we restrict ourselves to linear models. In CRR, the ones proposed are in first differences, because the unit root tests do not reject the null of one unit root. Hence, they try the univariate models

$$\Delta\mathrm{cp}_t = \varepsilon_t; \quad \Delta\mathrm{cp}_t = \beta_0 + \varepsilon_t; \quad \Delta\mathrm{cp}_t = \beta_0 + \beta_1\Delta\mathrm{cp}_{t-1} + \varepsilon_t,$$

where ε_t is white noise. These models are a random walk (RW), a random walk with drift (RWD) and an AR(1) (in fact, referred to the original variable cp_t it is an ARI(1,1), but we respect the nomenclature by CRR). The bivariate models are

$$\Delta\mathrm{cp}_t = \beta_0 + \gamma\Delta\mathrm{er}_{t-1} + \varepsilon_t; \quad \Delta\mathrm{cp}_t = \beta_0 + \beta_1\Delta\mathrm{cp}_{t-1} + \gamma\Delta\mathrm{er}_{t-1} + \varepsilon_t,$$

that is, a simple regression model (R) and a VAR(1). Here by er_t we mean generically, any of the three exchange rates.

With these models, different specific null hypotheses can be tested: (i) R does not beat RW, (ii) R does not beat RWD and (iii) VAR(1) does not beat AR(1)[2]. Firstly, these nulls are tested with in-sample Granger-causality tests, that consist on testing the significance of the parameters of er_{t-1} in the bivariate

[2] In fact, it is also possible to compare RW and RWD to the VAR(1), but in the event of rejection we would not know whether we do because of $\Delta\mathrm{cp}_{t-1}$ or $\Delta\mathrm{er}_{t-1}$.

models. To reject the null amounts to say the bivariate model is better than the univariate one. To account for possible instabilities of the parameters, the procedure[3] of Rossi (2005) is also used. These tests support the idea that the exchange rates actually help forecasting $\mathrm{cp}_{t+\tau}$.

However, there is a widespread belief that in-sample (IS) tests tend to reject too often and thus they are unreliable when they indicate Granger causality. This has been contested, for example in Inoue and Kilian (2005). But however compelling are their arguments, out-of-sample (OOS) tests are still favored by practitioners and in fact, the literature about them is still growing. There are strong intuitive reasons why OOS tests are more convincing than IS tests. When using IS tests, rejection of the null of no GC can be regarded as evaluating the models using forecast errors in which the forecasts use the future (because future observations are used to estimate the parameters). With OOS tests, we use true or 'pure' forecasts, in that they do not use the future[4].

4.2 Pairwise tests

For OOS causality testing, it is used the ENC-NEW test from Clark and McCracken (2001). This test is subsequently called in the literature ENC-F, so we stick to this notation. To apply this kind of tests, the series are split into two parts of lengths R and P. Then, there are different ways to proceed. The one used here is called 'rolling scheme' and it consists of estimating the parameters with a sliding window that comprises observations $t - R + 1$ to t, with $t = R, \dots, R + P$. The 'recursive scheme' uses an expanding window from 1 to t and the 'fixed scheme' keeps the first estimates throughout.

Encompassing tests are specifically designed to deal with nested comparisons. When the nested model is in population terms as good as the larger one,

[3] Which essentially amounts to split the sample in two parts, enlarge the ordinary Granger-Causality test with the difference between the estimates in both parts and average along the splitting points.

[4] With a small caveat: they are not absolutely realistic because use data from later vintages to forecast a certain period.

the finite-sample MSFE of the larger model will be greater due to estimation variability. These tests get more power by correcting for this effect. The statistic of the ENC-F test is

$$\text{ENC-F} = P\frac{P^{-1}\sum_{t=R+1}^{R+P}(\hat{u}_{2,t}^2 - \hat{u}_{2,t}\hat{u}_{1,t})}{P^{-1}\sum_{t=R+1}^{R+P}\hat{u}_{1,t}^2},$$

where $\hat{u}_{1,t}, \hat{u}_{2,t}$ are the forecasting errors of the nesting and nested model respectively (the indexes are reversed with respect to the original notation to fit with the MATLAB code below).

The asymptotic distribution of the test is non-standard and depends on the estimation scheme and on the limit of the ratio R/P, but critical values can be found for many situations in the appendix to Clark and McCracken (2001). They also can be obtained by a semiparametric bootstrap (Clark and McCracken, 2012).

In CRR, three tests are performed: RW vs R, RWD vs R and AR(1) vs VAR(1). The results reported in tables 4, 7(a) panel D and 7(b) panel D of CRR, seem to point that exchange rates actually have predictive content to forecast cp_t. Unfortunately, their figures are affected by a MATLAB programming bug. The ENC-F statistic is computed for the rolling scheme in the function `testshacAR` as

```
Pred*(media(u2.^2-u1roll.*u2))/(media((u1roll-media(u1roll)).^2));
```

First of all, the denominator is not the MSFE, but the variance, which given that the nesting models have an intercept probably is asymptotically correct, but it is not the definition of the test. However, the main problem is that `u2` is the error of the nested model with the recursive scheme, so the formula is mixing different schemes and thus, the asymptotic critical values of Clark and McCracken (2001) are no longer valid.

We have run the right tests and a very different picture emerges in table I. We find only signs of a relevant improvement using exchange rates when the univariate model is the RW (and in the case of NZ, the RWD). In view of this, one may suspect that the rejections are mostly due to the poor forecasting

Table I

	AUS	NZ	CAN	CHI	SA
			Dollar ER		
RW vs R	4.7349(**)	6.0469(***)	0.2453	1.7840(*)	5.4078(***)
RWD vs R	0.6605	2.7718(**)	-0.5683	0.9560	-0.9793
AR(1) vs VAR(1)	-1.0233	0.8956	-0.6125	0.4096	-1.2851
RW vs R	0.0043	0.0026	0.3161	0.1121	0.0046
RWD vs R	0.1599	0.0217	0.8197	0.1433	0.8538
AR(1) vs VAR(1)	0.9820	0.1353	0.8876	0.2408	0.9434
			Nominal Effective ER		
RW vs R	4.5707(**)	4.2938(**)	0.4624	0.1356(*)	5.9648(***)
RWD vs R	0.9011	1.1783	-0.4594	0.1784	-0.0470
AR(1) vs VAR(1)	0.0694	-0.1482	-0.5093	-0.1374	-0.1066
RW vs R	0.0034	0.0116	0.2814	0.3577	0.0025
RWD vs R	0.0774	0.0784	0.6547	0.2978	0.4610
AR(1) vs VAR(1)	0.3741	0.4869	0.6665	0.4476	0.4798
			British pound ER		
RW vs R	3.5920(**)	2.5224(**)	0.3665	1.4283(*)	6.0028(***)
RWD vs R	-0.3618	-0.3217	-0.5013	0.7782	-0.5437
AR(1) vs VAR(1)	-0.2135	-2.0156	-0.6172	0.8781	-0.7954
RW vs R	0.0092	0.0383	0.2833	0.1237	0.0024
RWD vs R	0.8591	0.6444	0.8489	0.1258	0.7722
AR(1) vs VAR(1)	0.6631	1.0000	0.9191	0.1301	0.9157

In the first, third, and fifth panels, we report the ENC-F statistics of the pairwise comparisons. The asterisk indicate significance at 90% (*), 95% (**) and 99% (***) according to the critical values in Clark and McCracken (2001). The second, fourth, and sixth panels includes the p-values according to the bootstrap, that largely coincide.

Table II

	AUS	NZ	CAN	CHI	SA
RW vs AR(1)	22.4012(***)	43.1662(***)	-0.8938	2.5620(*)	5.9473(***)
RWD vs AR(1)	14.7158(***)	33.5563(***)	-1.7126	1.9542(**)	-0.4717

Comparison between univariate models. In the above panel, we report the ENC-F statistics of the pairwise comparisons. The asterisk indicate significance at 90% (*), 95% (**) and 99% (***) according to the critical values in Clark and McCracken (2001).

performance of the RW and RWD when compared to the AR(1). We can see this by performing encompassing tests to check whether the AR(1) are significantly better than the RW and RWD. In table II we present the results of the tests and we see that with the exception of Canada, the RW is inferior to the AR(1). Canada is precisely the country in which even with the RW, there is no rejection. On the other hand, the RWD is inferior to AR(1) for NZ.

This already points to the conclusion that OOS forecasting gives no evidence that exchange rates have actual predictive content. However, in the next section, we will discuss how can we deal more systematically with this kind of multiple comparisons.

4.3 Results with the multi-model RC

The situation we have here is that we want to test the null that the best univariate model of cp_t is as good as any bivariate one that uses past values of er_t. To find a feasible testing procedure, we first make the simplifying assumption that only linear models matter. Then, even among linear models, we have to restrict ourselves to relatively simple models. In CRR, this constraint is tight because using only pairwise comparisons with the ENC-F test, it would be difficult to manage a large number of models. Even with so few models, pairwise comparisons may produce contradictory results, as we saw in the previous section. How

Table III

	AUS	NZ	CAN	CHI	SA
			Dollar ER		
Δ-Mm	0.9810	0.1343	0.8526	0.2102	0.8413
Δ-mM	0.9817	0.1346	0.8654	0.2150	0.8499
			Nominal Effective ER		
Δ-Mm	0.3782	0.4914	0.5862	0.3966	0.3881
Δ-mM	0.4027	0.5476	0.6183	0.4162	0.3960
			British pound ER		
Δ-Mm	0.6629	0.9998	0.8206	0.1043	0.7311
Δ-mM	0.6800	0.9998	0.8484	0.1074	0.7454

P-values of the maximin and minimax tests for the three settings.

can we summarize the information into a simple decision?

Let us first consider the case that we have only one univariate model, say $\Delta \mathrm{cp}_{t+1} = \beta_0 + \varepsilon_{t+1}$ and we want to compare it with different alternatives, for example, $\Delta \mathrm{cp}_{t+1} = \beta_0 + \gamma_1 \Delta \mathrm{er}_t + \varepsilon_{t+1}$ and $\Delta \mathrm{cp}_{t+1} = \beta_0 + \gamma_1 \Delta \mathrm{er}_t + \gamma_2 \Delta \mathrm{er}_{t-1} + \varepsilon_{t+1}$. In this situation, we can use an RC.

However, RC tests still require to use a particular benchmark. In the case of CRR, it is not clear beforehand which benchmark should we use. We can deal with this in different ways: either using some model selection criteria to decide the benchmark model or generalizing the RC test.

In table III, we report the p-values of the maximin and minimax tests applied to the benchmarks RW, RWD and AR(1) and the alternatives R and VAR(1). It seems that when we take into account all models, there is no evidence that bivariate models forecast the commodity prices better than the best univariate model.

Given that we have now a tool to combine the information the pairwise comparisons between two classes of models, we do not need to constrain ourselves to so simple models. We can include regressions with more lags of both variables. We will assume now that I contains univariate models such as

$$\Delta \mathrm{cp}_t = \phi_1 \Delta \mathrm{cp}_{t-1} + \ldots + \phi_p \Delta \mathrm{cp}_{t-p} + \varepsilon_t,$$

and J bivariate models as

$$\begin{aligned} \Delta \mathrm{cp}_t = {} & \phi_{11} \Delta \mathrm{cp}_{t-1} + \ldots + \phi_{1p} \Delta \mathrm{cp}_{t-p} + \\ & \phi_{21} \Delta \mathrm{er}_{t-1} + \ldots + \phi_{2p} \Delta \mathrm{er}_{t-p} + \varepsilon_t, \end{aligned}$$

with p up to 12. For the data of South Africa, the largest models have multicollinearity troubles, so we reduce the order of the models to 9. The p-values of the tests with these sets of models are reported in table IV. Here it seems that some signs of Granger-causality appear in the Australian data. However, in order not to incur in data mining, we should take into account that even if the null holds for the five countries, the probability of one of them having a p-value under 0.1 is over 40%, so we should be cautious about giving relevance to this, especially given that the p-value is only slightly below 0.1.

On the other hand, in view that method 2 appears to perform well, we could just look in table I at the comparisons or AR(1) against VAR(1), but this is consistent with the minimax and maximin tests.

The Δ-Mm and Δ-mM tests can be applied to longer horizons. In table V, we report the results of the tests with $\tau = 4$. Again, we do not find evidence of predictability.

These results may appear to contradict the findings of Clark and McCracken (2012). However, in their experiment, the alternative models use not just exchange rates, but also a commodity futures index. Table 5 in their article presents the alternative models sorted by their RMSE. We can see that most of the models that perform well include the futures index and thus, the models that use only exchange rates are mostly in the bottom of the table. Hence, we cannot conclude that rejection is due to the use of exchange rates instead of the future indexes.

Table IV

	AUS	NZ	CAN	CHI	SA[†]
			Dollar ER		
Δ-Mm	0.0626	0.7666	0.6762	0.3176	0.2469
Δ-mM	0.0814	0.8506	0.6412	0.3023	0.2859
			Nominal Effective ER		
Δ-Mm	0.0949	0.1399	0.9068	0.6493	0.2334
Δ-mM	0.1215	0.1488	0.7628	0.7931	0.2294
			British pound ER		
Δ-Mm	0.1399	0.2086	0.7489	0.2139	0.3066
Δ-mM	0.1662	0.2350	0.6943	0.2487	0.3749

These are p-values of the tests with autoregressive models of order up to 12. [†] For SA, maximum lag is set to 9, because of multicollinearity problems that appear for larger models.

Table V

	AUS	NZ	CAN	CHI	SA
			Dollar ER		
Δ-Mm	0.3353	0.1812	0.2194	0.4888	0.7976
Δ-mM	0.3428	0.1441	0.2237	0.4981	0.8081
			Nominal Effective ER		
Δ-Mm	0.5940	0.5667	0.5097	0.8731	0.3556
Δ-mM	0.6153	0.5728	0.5304	0.8799	0.3283
			British pound ER		
Δ-Mm	0.5800	0.4468	0.8755	0.2238	0.5901
Δ-mM	0.5965	0.4540	0.8783	0.2323	0.5655

P-values of the maximin and minimax tests for the three settings, with forecasting horizon $\tau = 4$.

5 Final remarks

We have generalized the idea of RC to the case in which one wants to compare two classes of models. This makes unnecessary to pick a specific benchmark and apply an RC, which is convenient because we have seen that by being too cavalier about how to choose the benchmark we can easily get into troubles.

In order to obtain critical values for the test, we use a semiparametric bootstrap. The assumptions necessary for the bootstrap to work properly are the main cause of lack of generality. Hence, a theme for future research would be to find a more general way to approximate or simulate the distribution of the test. Nevertheless, even in this constrained framework there are a number of examples in which the test may be of interest.

In the past, it had been difficult to find empirical evidence for models relating exchange rates to its fundamentals. Unfortunately, our results seem to indicate that the commodity price approach is no less hard and it resist our efforts to find predictability.

A Proofs and lemmas

Proof of Proposition 1. We will prove (i) for the MSE-t-mM and leave the other case to the reader. First

$$\Delta\text{-mM} = \min\big(\min_{i\in I_0}\max_{j\in J}\Delta_{ij}, \min_{i\in I\setminus I_0}\max_{j\in J}\Delta_{ij}\big).$$

Since $\forall i \in I \setminus I_0, \max_{j\in J}\Delta_{ij} \xrightarrow{p} \infty$, then $\min_{i\in I\setminus I_0}\max_{j\in J}\Delta_{ij} \xrightarrow{p} \infty$. Thus, by lemma 1,

$$\Delta\text{-mM} = \min_{i\in I_0}\max_{j\in J}\Delta_{ij} + o_p(1). \tag{7}$$

Now, we can apply again lemma 1 for each i to get $\max_{j\in J}\Delta_{ij} = \max_{j\in J_0}\Delta_{ij} + o_p(1)$. Then, we conclude by replacing in (7) and invoking the continuous mapping theorem. □

Lemma 2. *Let the regressors be ordered so that $\beta = (\beta_\ell', \beta_{\ell^c}')'$. Then, if $\forall a \in \{1,\ldots,n\} \setminus \ell, \sigma^2(\ell) = \sigma^2(\ell \cap \{a\})$, then $\beta_{\ell^c} = 0$.*

Proof. Let us assume for simplicity that $\mathbb{E}x_t = 0$. Now, $B_{\mu,\nu} = \mathbb{E}x_{\mu,t}x'_{\nu,t}$ and $b_\mu = \mathbb{E}x_{\mu,t}y'_{t+\tau}$. Then, the assumption implies that

$$\begin{pmatrix} B_{\ell,\ell} & B_{\ell,a} \\ B_{a,\ell} & B_{a,a} \end{pmatrix}^{-1} \begin{pmatrix} b_\ell \\ b_a \end{pmatrix} = \begin{pmatrix} \nu_a \\ 0 \end{pmatrix}.$$

Hence, $B_{\ell,\ell} = \nu b_\ell$ and $B_{\ell,a} = \nu_a b_a$, so all ν_a are equal to ν. This implies that $\pi = (\nu', 0, \dots, 0)'$ is a solution to the system

$$\begin{pmatrix} B_{\ell,\ell} & B_{\ell,\ell^c} \\ B_{\ell^c,\ell} & B_{\ell^c,\ell^c} \end{pmatrix} \pi = \begin{pmatrix} b_\ell \\ b_{\ell^c} \end{pmatrix}.$$

□

Proof of Proposition 2. Let us assume that $i \cup j = \{1, \dots n\}$ and $\ell = i$. Now, if i and j were in I_0 and J_0 respectively, then $\sigma^2(i \cup \{a\}) = \sigma^2(i)$ for all $a \in j \setminus i$. Then, by lemma 2 the coefficients in the regression $y_{t+\tau}$ on $x_{i\cup j,t}$ of all the elements of $j \setminus i$ would be zero. This contradicts the assumption that i and j were nonnested. □

Proof of Proposition 3. The result is a straightforward adaptation of theorem 3.2 in Clark and McCracken (2012) for the pairs (i, j) where i is nested in j and theorem 2.1 in Clark and McCracken (2014) when i and j are overlapping. We only need to be careful to ensure that the Brownian motion used to express the distribution of the one-to-one statistic is the same for all pairs (i, j). For this, note that if we set $Ck_t = h_t$, where $C = S_{hh}^{1/2}$, we can apply theorem 4.1 in Hansen (1992) to get

$$\sum_{s=[ut]}^{[vt]} K(s)k'_s \Rightarrow \int_u^v W(\omega)dW(\omega)',$$

where $\Rightarrow$ denotes weak convergence and $K(s) = C^{-1}H(s)$. Now, $\tilde{h}_s = \sigma^{-1}\tilde{A}B^{1/2}Ck_s$. Thus

$$\sum_{s=[ut]}^{t} \tilde{H}(s)\tilde{h}'_s = \sigma^{-2}\tilde{A}B^{1/2}C\sum_{s}^{t} K(t)k'_sC'B^{1/2\prime}\tilde{A}' \Rightarrow$$

$$\sigma^{-2}\tilde{A}B^{1/2}C\Big\{\int_u^1 W(\omega)dW(\omega)'\Big\}C'B^{1/2\prime}\tilde{A}'.$$

Consequently,

$$\sum_{s=[ut]}^{t} \tilde{h}_s' \tilde{H}(s) = \text{tr} \sum_{s=[ut]}^{t} \tilde{H}(s)\tilde{h}_s' \Rightarrow$$

$$\sigma^{-2}\text{tr}\tilde{A}B^{1/2}C\Big\{\int_u^1 W(\omega)dW(\omega)'\Big\}C'B^{1/2'}\tilde{A}' =$$

$$\int_u^1 W(\omega)'\Big(\sigma^{-2}C'B^{1/2'}\tilde{A}'\tilde{A}B^{1/2}C\Big)dW(\omega).$$

The sums for $\Gamma_{2,ij}$ and $\Gamma_{3,ij}$ can be dealt with in the same way. □

Proof of Proposition 5. We will prove (iii) first. Lemma 2 and assumption $2'$(c) imply that $\beta_{k_0^c} = 0$. On the other hand, assumption $2'$(c) ensures that all models $i \in I_0$ and $j \in J_0$ satisfy $k_1 \subset i, j$, where k_1 is the set of the regressors with nonzero coefficients in the full model. Now, let us compare our bootstrap with those of Clark and McCracken (2012) and (2014) when $(i, j) \in I_0 \times J_0$.

- In case $i \in I_0$ is nested in $j \in J_0$, the difference is that we generate the data with $x_{k_0,t}'\hat{\beta}_{k_0,T} + \hat{v}_{t+\tau}^*$ instead of $x_{i,t}'\beta_i + \hat{v}_{t+\tau}^*$, but since all the regressors with nonzero coefficients are in $i \subset k_0$, then the excess parameters have null population values.

- In case $i \in I_0$ and $j \in J_0$ are overlapping, the difference is that we generate the data with $x_{k_0,t}'\hat{\beta}_{k_0,T} + \hat{v}_{t+\tau}^*$ instead of $x_{i\cap j,t}'\beta_{i\cap j} + \hat{v}_{t+\tau}^*$, but now, all regressors with nonzero coefficients are in $i \cap j \subset k_0$ and again, the excess parameters have null population values.

We have to see that the excess regressors in the artificial sample only aport a $o_{p^*}(1)$ error to the MSE-F statistics. Let us check this for the numerator. For the denominators of both pairwise statistics, the calculations are similar.

We denote by $\hat{u}_{\ell,t+\tau}^*(a)$ the bootstrapped residual obtained using the model $x_{a,t}'\hat{\beta}_{a,T} + \hat{v}_{t+\tau}^*$. We can prove

$$\sum_s \Big\{(\hat{u}_{i,s+\tau}^*(i)^2 - \hat{u}_{j,s+\tau}^*(i)^2) - (\hat{u}_{i,s+\tau}^*(k_0)^2 - \hat{u}_{j,s+\tau}^*(k_0)^2)\Big\} = o_p(1). \quad (8)$$

In order to see that (8) holds, we write first, for $\ell = i, j$,

$$\hat{u}_{\ell,s+\tau}^*(k_0) = \hat{u}_{\ell,s+\tau}^*(i) + x_t' Q_j(t)B(t)^{-1}J_{k_0}\hat{\beta}_{k_0,T}.$$

On the other hand, we can put $J_{k_0}\hat{\beta}_{k_0,T} = J_{k_0\cap\ell}J'_{k_0\cap\ell}J_{k_0}\hat{\beta}_{k_0,T} + J_{k_0\cap\ell^c}J'_{k_0\cap\ell^c}J_{k_0}\hat{\beta}_{k_0,T}$, but $\forall m \subset \ell, Q_\ell(t)B(t)^{-1}J_m = 0$ and thus,

$$\hat{u}^*_{\ell,s+\tau}(k_0) = \hat{u}^*_{\ell,s+\tau}(i) + x'_t Q_\ell(t)B(t)^{-1}J_{k_0\cap\ell^c}J'_{k_0\cap\ell^c}J_{k_0}\hat{\beta}_{k_0,T} = \hat{u}^*_{\ell,s+\tau}(i) + D_\ell.$$

Hence,

$$\hat{u}^*_{i,s+\tau}(k_0)^2 - \hat{u}^*_{j,s+\tau}(k_0)^2 = \hat{u}^*_{i,s+\tau}(i)^2 - \hat{u}^*_{j,s+\tau}(i)^2 +$$
$$2\hat{u}^*_{i,s+\tau}(i)(D_i - D_j) + 2D_j\big(\hat{u}^*_{i,s+\tau}(i) - \hat{u}^*_{j,s+\tau}(i)\big) + (D_i^2 - D_j^2) =$$
$$\hat{u}^*_{i,s+\tau}(i)^2 - \hat{u}^*_{j,s+\tau}(i)^2 + E_1 + E_2 + E_3.$$

Now, that

$$E_1 = \sum_s 2\hat{h}^*_{i,s+\tau}(i)'\big(Q_i(s) - Q_j(s)\big)B(t)^{-1} \times$$
$$\Big[J_{k_0\cap i^c}J'_{k_0\cap i^c} - J_{k_0\cap j^c}J'_{k_0\cap j^c}\Big]J_{k_0}\hat{\beta}_{k_0,T} = o_p(1)$$

can be proved along the lines of lemma 1 in Clark and McCracken (2012), and using that $[J_{k_0\cap i^c}J'_{k_0\cap i^c} - J_{k_0\cap j^c}J'_{k_0\cap j^c}]J_{k_0}\hat{\beta}_{k_0,T} = o_p(T^{-1/2})$. Now,

$$|E_2| = |\sum_s \hat{H}^*(s)'(Q_i(s) - Q_j(s))x_s x'_s Q_i(t)B(t)^{-1}J_{k_0\cap i^c}J'_{k_0\cap i^c}J_{k_0}\hat{\beta}_{k_0,T}| \leq$$
$$\sup_s |\hat{H}^*(s)| \cdot \sum_s |(Q_i(s) - Q_j(s))x_s x'_s Q_i(t)B(t)^{-1}| \cdot |J_{k_0\cap i^c}J'_{k_0\cap i^c}J_{k_0}\hat{\beta}_{k_0,T}|.$$

The first factor is $O_{p^*}(T^{-1/2})$, the second is $O_{p^*}(T)$ and the third is $o_{p^*}(T^{-1/2})$ because the population parameters outside i are zero. For

$$E_3 = \sum_s \hat{\beta}'_{k_0,T}J'_{k_0}J_{k_0\cap i^c}J'_{k_0\cap i^c}B(s)^{-1}Q_i(s)x_s x'_s(Q_i(s) - Q_j(s)) \times$$
$$B(s)^{-1}J_{k_0\cap i^c}J'_{k_0\cap i^c}J_{k_0}\hat{\beta}_{k_0,T} = o_p(1),$$

we may use that $\sup_s |T^{1/2}(Q_i(s) - Q_j(s) - Q_i + Q_j)| = O_p(1)$.

For (i) and (ii), we use that by lemma 1 in Clark and McCracken (2012) and (8), $\hat{\sigma}^{2,*}_\tau(\ell) \xrightarrow{p^*} \sigma^2_\tau(\ell)$, that is, the bootstrapped MSE are consistent. On the other hand, the same arguments can be used to prove that the bootstrap estimates of the autocovariance $\hat{\gamma}^*_{d_{ij}}(l)$ are consistent.

Therefore, when $(i,j) \in I_0 \times J_0^c$, both $\hat{\sigma}_\tau^{2,*}(i) - \hat{\sigma}_\tau^{2,*}(j) \xrightarrow{p^*} \sigma_\tau^2(i) - \sigma_\tau^2(j) < 0$ and the numerator of MSE-t_{ij}^* converges to a positive variance. Thus, MSE-$\text{t}_{ij}^* \to -\infty$. Conversely, when $(i,j) \in I_0^c \times J_0$, MSE-$\text{t}_{ij}^* \to +\infty$. The same arguments used in the proof of proposition 1 entail that only the models in I_0 and J_0 are asymptotically relevant. □

References

[1] Chen, Yu-Chin, Kenneth S. Rogoff, and Barbara Rossi, "Can exchange rates forecast commodity prices?," *The Quarterly Journal of Economics*, 125 (2010), 1145-1194.

[2] Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual, "Empirical exchange rate models of the nineties: Are any fit to survive?," *Journal of International Money and Finance*, 24 (2005), 1150-1175.

[3] Clark, Todd E., and Michael W. McCracken, "Tests of equal forecast accuracy and encompassing for nested models," *Journal of econometrics*, 105 (2001), 85-110.

[4] Clark, Todd E., and Michael W. McCracken, "Evaluating direct multistep forecasts," *Econometric Reviews*, 24 (2005), 369-404.

[5] Clark, Todd E., and Michael W. McCracken, "Reality checks and comparisons of nested predictive models," *Journal of Business and Economic Statistics*, 30 (2012) 53-66.

[6] Clark, Todd E., and Michael W. McCracken, "Tests of equal forecast accuracy for overlapping models, *Journal of Applied Econometrics* 29 (2014), 415-430.

[7] Davidson, Russell, and Emmanuel Flachaire, "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146 (2008), 162-169.

[8] Engel, Charles, and Kenneth D. West, "Exchange rates and fundamentals," NBER Working Paper No. w10723, 2004.

[9] Giacomini, Raffaella, and Halbert White, "Tests of conditional predictive ability," *Econometrica*, 74 (2006), 1545-1578.

[10] Hansen, Peter Reinhard, "A test for superior predictive ability," *Journal of Business and Economic Statistics*, 23 (2005) 365-380.

[11] Hansen, Bruce E. "Convergence to stochastic integrals for dependent heterogeneous processes," *Econometric Theory*, 8 (1992), 489-500.

[12] Inoue, Atsushi, and Lutz Kilian, "In-sample or out-of-sample tests of predictability: Which one should we use?," *Econometric Reviews*, 23 (2005) 371-402.

[13] Matilla-García, Mariano, Ruiz Marín, Manuel and Dore, Mohammed I., "A permutation entropy based test for causality: The volume–stock price relation," *Physica A*, 398 (2014), 280-288.

[14] McCracken, Michael W. "Parameter estimation and tests of equal forecast accuracy between non-nested models," *International Journal of Forecasting* 20 (2004), 503-514.

[15] Meese, Richard A., and Kenneth Rogoff, "Empirical exchange rate models of the seventies: do they fit out of sample?," *Journal of international economics*, 14 (1983), 3-24.

[16] Meese, Richard, and Kenneth Rogoff, "The out-of-sample failure of empirical exchange rate models: sampling error or misspecification?," in *Exchange rates and international macroeconomics*, Jacob A. Frenkel, ed. (University of Chicago Press, 1983).

[17] Politis, Dimitris N., and Joseph P. Romano, "The stationary bootstrap," *Journal of the American Statistical Association*, 89 (1994), 1303-1313.

[18] Rossi, Barbara. "Optimal tests for nested model selection with underlying parameter instability," *Econometric theory*, 21 (2005), 962-990.

[19] Vuong, Q. H., 1989, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 307-333.

[20] West, Kenneth. D. Asymptotic inference about predictive ability. Econometrica: Journal of the Econometric Society 64 (1996), 1067-1084.

[21] White, Halbert, "A reality check for data snooping," *Econometrica*, 68 (2000), 1097-1126.