

Reality checks with multiple benchmarks*

Ignacio Arbués,[†] Cristina Casaseca, Ramiro Ledo and Silvia Rama
Instituto Nacional de Estadística

March 11, 2013

Abstract

Reality Checks (RC) are tests designed to deal with situations in which there is a, possibly large, class of competing forecasting models and one wants to know if any of these models beats a given benchmark.

However, there are situations in which it is more desirable to allow for more than one benchmark. For example, we may want to test Granger causality by comparing two classes of models distinguished by the inclusion of a certain regressor.

Thus, we propose to test the null that none of the alternative models beats the best of a set of benchmarks. This null can be tested by a procedure that consists in a straightforward generalization of the statistic of one of the Reality Checks for nested models of Clark and McCracken (2012) and a bootstrap that generates the artificial samples according to a model that nests all the benchmarks. We show by simulations that the test has the correct empirical size and its power is greater than just applying the RC test to the best performing benchmark. An application to causality is presented as well.

*We want to express our gratitude to an anonymous referee whose advice has been conducive to a great improvement of our work.

[†]Corresponding author. Address: Instituto Nacional de Estadística. Castellana, 183, 28071, Madrid, Spain. E-mail: iarbues@ine.es. Telephone: +34 915834641.

1 Introduction

The Reality Check (RC) was proposed by White (2000) as a means to deal with situations in which several forecasting models are available. Suppose that in a certain application, some of them produce better out-of-sample forecasts than a given benchmark (usually a simpler model). If the number of models is large, this may well happen by chance and not because there is actually a model better than the benchmark. White proposed a test for the null that the benchmark 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. As a consequence of the asymptotic theory of 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 (SPA) test (Hansen, 2005) is a modification of White's Reality Check in which greater power is obtained by normalizing the mean differences by estimates of their standard deviations.

A problem shared by 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 RC and Hansen's SPA only holds in the nested case when the out-of-sample size P is small compared to the in-sample size T (more precisely, when the ratio P/T converges to zero).

The theory of predictive ability tests for nested models started shortly after Diebold and Mariano (1995) and West (1996) introduced the kind of predictive ability test that is usually known as the Diebold and Mariano test. It became apparent that for nested models, the tests were not asymptotically normal and the critical values would have to be drawn from non-standard distributions that can be expressed as functionals of Brownian motions. The asymptotic theory of these tests for the nested, one-step forecast case was developed in Clark and McCracken (2001) and McCracken (2004). On the other hand, in Clark and

West (2007), it is argued that the normal distribution can be an acceptable approximation in some cases. 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 Reality Check converge 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. In this article, we propose a generalization of this RC. In particular, we are interested in situations where there is not a unique benchmark model, but a set of possible ones. In fact, the case in which the benchmark is perfectly determined beforehand is not the most natural in macro-economic data (it may be, however, in financial examples such as the one in White (2000), where the natural benchmark is that the predictand is a random walk). We can also describe the test as a comparison between two classes of models in which the null is that the best of one class is at least as good as the best of the other one.

We had in mind one particular application, namely, to determine if a certain variable has predictive capacity in forecasting another one or, in other words, to test Granger causality. This idea underlies some applications, such as the one in Hansen (2005). We will show in section 6 that the result of the test can be strongly dependent on the choice of the benchmark. If the benchmark is not good, the result of the test can mislead to the conclusion that there is a causality relation when in fact, there is not.

Other examples of such comparisons between forecasting model classes are those mentioned in Pincheira (2011): (1) time-series models compared to economic models, (2) simple combination strategies vs complex combination schemes, (3) models that use the aggregate CPI and models that use disaggregate components and (4) linear and nonlinear models.

Consequently, we want a test for the null that the best of a class of benchmarks is at least as good as the best of a class of alternatives. There are at least two obvious ways to generalize the MSE-t statistic of Clark and McCracken

(2012) to the case of multiple benchmarks and we analyze both of them. We obtain the asymptotic distribution of the test and propose a variant of their fixed-regressor wild bootstrap for this case. We also show the results of some Monte Carlo experiments that are designed to resemble some cases of interest and an application to real data.

In section 2 we introduce the less technical assumptions, the tests and the essential notation. In section 3 we describe the asymptotic behavior of the statistics, relegating the details to a mathematical appendix. In section 4, we propose a variation of the wild bootstrap. Finally, the results of the simulations (section 5) and of the real-data application (section 6) are presented.

2 Comparison of multiple benchmarks to multiple alternatives

In this section, we will begin by motivating the test with a theoretical example and describing the environment, notation and the assumptions that are required for its contents. The most technical assumptions are confined to the mathematical appendix and here we only discuss those that are necessary to understand the implications of the null hypothesis.

2.1 A theoretical example

Assume that the bivariate series $(x_t^{(1)}, y_t)$ satisfies the model

$$\begin{aligned}x_t^{(1)} &= \phi_1 y_{t-1} + \varepsilon_{1,t}, \\y_t &= \phi_2 y_{t-2} + \varepsilon_{2,t},\end{aligned}$$

where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are uncorrelated. Then, $x_t^{(1)}$ does not Granger-cause y_t and the optimal predictor of y_{t+1} using $\{x_{t-j}, y_{t-j}\}_{j=0}^p$ is $\hat{y}_{t+1} = \phi_2 y_{t-1}$ for any $p \geq 1$. However, for $p = 0$, there is a predictor of the form $\hat{y}_{t+1} = \beta_2 x_t^{(1)}$, with $\beta_2 \neq 0$ that has less Mean Squared Error (MSE) than the optimal univariate AR(1) predictor, that is $\hat{y}_{t+1} = 0$. Suppose that we are performing a RC test to see if it is possible to predict better y_{t+1} using as inputs some of the indicators $\{x_t^{(k)}\}_k$. If the benchmark is the AR(1), then we will probably reject because

of $x_t^{(1)}$, when in fact, the AR(2) is the best model and it cannot be improved using $x_t^{(1)}$.

On the other hand, if we design a test for the null that no alternative model beats the best AR(p) with $p \leq P$, we can avoid that kind of errors, or at least to make them less likely when P is large enough.

2.2 Environment and null hypothesis

The notation will be quite similar to Clark and McCracken (2012). We consider forecasts of $y_{t+\tau}$ using linear models. By n , we denote the number of available predictors. We will use indexes such as i to denote subsets of $\{1, \dots, n\}$ and $x_{i,t}$ is the vector that comprises the values of the predictors that belong to the set i at time t . The indexes i can be equivalently interpreted as models. The vector that includes all the n predictors is denoted by x_t .

Assumption 1. *For any i , we estimate the vector of parameters $\hat{\beta}_{i,t}$ by direct linear squares with a recursive scheme, that is*

$$\hat{\beta}_{i,t} = \arg \min_{\beta} \sum_{s=1}^t |y_{s+\tau} - x'_{i,s} \beta|^2.$$

With the estimated parameters, we build forecasts $\hat{y}_{i,t+\tau} = x'_{i,t} \hat{\beta}_{i,t}$ for $t = T, \dots, T + P - \tau$. 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 . 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 .

We are interested in the mean squared forecasting errors. Let $\sigma_{\tau}^2(\ell)$ be the population mean squared error 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 $u_t = y_{t+\tau} - x'_t \beta^*$.

The null hypothesis is

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

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

We want to preserve one idea of the single-benchmark RC, namely, that the benchmark is a simpler model. However, it would be too strong to assume that

all $i \in I$ are nested in all $j \in J$. For example, when testing causality of a variable z_t to y_t , I should contain models with only lags of y_t , whereas the models in J would 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, what is clearly undesirable.

The following assumption gives space for many applications, as we will see in section 5 while preserving the idea of small vs large models, simplifying the asymptotic theory and allowing to obtain the critical values by a fixed-regressor bootstrap.

Assumption 2. (a) $\forall i \in I$ and $a \in \{1, \dots, n\}$, $\exists j \in J$ such that $i \cup \{a\} \subseteq j$.
(b) There is a certain $k_0 \subset \{1, \dots, n\}$ such that $\forall i \in I, i \subseteq k_0$.

In the case of Granger causality, k_0 is the set of the lags of the predictand and thus, $i \in I$ represent univariate models and $j \in J$ multivariate ones.

2.3 Test statistics

Let $\hat{d}_{ij,t} = \hat{u}_{i,t+\tau}^2 - \hat{u}_{j,t+\tau}^2$, $\bar{d}_{ij} = (P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} \hat{d}_{ij,t}$ and $\hat{\gamma}_{d_{ij}}(l) = (P - \tau + 1)^{-1} \sum_{t=T+l}^{T+P-\tau} (\hat{d}_{ij,t} - \bar{d}_{ij})(\hat{d}_{ij,t-l} - \bar{d}_{ij})$. We can use the covariances $\hat{\gamma}_{d_{ij}}(l)$ to estimate the long-run covariance of \bar{d}_{ij} as $(P - \tau + 1)^{-1/2} \hat{S}_{d_{ij}} = \sum_{l=-\bar{l}}^{\bar{l}} K(l/L) \hat{\gamma}_{d_{ij}}(l)$, where $K(\cdot)$ is a certain kernel and L is a truncation parameter and $\hat{\gamma}_{d_{ij}}(-l) = \hat{\gamma}_{d_{ij}}(l)$. Then, we build the one-to-one comparison statistics as $\text{MSE-t}_{ij} = (P - \tau + 1)^{-1/2} \bar{d}_{ij} / \hat{S}_{d_{ij}}$. Finally, the test statistics are

$$\text{MSE-t-mM} = \min_{i \in I} \max_{j \in J} \text{MSE-t}_{ij} \quad (1)$$

and

$$\text{MSE-t-Mm} = \max_{j \in J} \min_{i \in I} \text{MSE-t}_{ij}. \quad (2)$$

When I has only one model, then MSE-t-mM and MSE-t-Mm reduce to the MSE-t statistic of Clark and McCracken (2012). On the other hand, note that if $\text{MSE}_\ell = (P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} \hat{u}_{\ell,t}^2$, then $\min_{i \in I} \max_{j \in J} (\text{MSE}_i - \text{MSE}_j) = \max_{j \in J} \min_{i \in I} (\text{MSE}_i - \text{MSE}_j) = \min_{i \in I} \text{MSE}_i - \min_{j \in J} \text{MSE}_j$. Hence, the fact that MSE-t-mM and MSE-t-Mm do not coincide is a consequence of dividing the MSE differences by the variance estimator $\hat{S}_{d_{ij}}$. The choice among MSE-t-mM and MSE-t-Mm is relevant in more than one respect. As we show below,

there is a significant difference of power, but MSE-t-Mm is also more related to an extension of the test that we will outline in section 4.

3 Asymptotics

Let us denote by k_1 the set of the regressors that have nonzero parameters in the full model $x_t'\beta^*$. The null hypothesis and assumption 2 together imply that there is a reordering of the regressors such that $x_t = (x'_{k_0,t}, x'_{k_0^c,t})'$ and $\beta^* = (\beta_{k_0}^* ', \mathbf{0}')$. Now, let I_0 and J_0 be the sets of models in I and J that nest $k_1 \cap k_0$ and k_1 respectively. That is, I_0 comprises the "good" models of I and J_0 the "good" models of J .

Under the null, $k_1 \subseteq k_0$, so $k_1 \cap k_0 = k_1$ and all the good models include k_1 . Hence, all of them have the same population forecasting error, whereas the remaining ones, that is, the "bad" ones in $(I \setminus I_0) \cup (J \setminus J_0)$ have greater error.

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 MSE-t $_{ij}$ with $i \in I_0$ and $j \in J_0$ are asymptotically relevant. In other words, only the good models matter asymptotically. More precisely, in the mathematical appendix, we prove that

$$\text{MSE-t-mM} = \min_{i \in I_0} \max_{j \in J_0} \text{MSE-t}_{ij} + o_p(1), \quad (3)$$

$$\text{MSE-t-Mm} = \max_{j \in J_0} \min_{i \in I_0} \text{MSE-t}_{ij} + o_p(1). \quad (4)$$

Another consequence of the null hypothesis is that among the pairs $(i, j) \in I_0 \times J_0$, we can find only two situations. Either i is nested in j or the models are overlapping in the sense of Vuong (1989), that is, both contain the true model k_1 plus terms that vanish for the population value of the parameters.

Hence, for $(i, j) \in I_0 \times J_0$, the asymptotics of MSE-t $_{ij}$, is given by the following known results.

- (a) When i and j are nested, the asymptotic distribution of MSE-t $_{ij}$ is given by theorem 3.2 in Clark and McCracken (2012).
- (b) When i and j are overlapping, the asymptotic distribution of MSE-t $_{ij}$ is given by theorem 2.1 in Clark and McCracken (2011).

In the mathematical appendix, we give the specific non-standard distributions for both cases.

4 The bootstrap

Although we know the asymptotic distribution of the test, we will use a bootstrap to obtain the critical values or p-values. Our version of the bootstrap is adapted from Clark and McCracken (2011 and 2012). They generate artificial samples according to

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

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. In our case, the null hypothesis implies that the true model is included in k_0 , so we will generate our artificial samples according to $y_t^* = \hat{\beta}_{k_0} x_{k_0,t} + \hat{v}_t^*$. 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 $t = 1, \dots, T + P - \tau$.
- 2 We estimate a MA($\tau - 1$) model $v_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_{\tau-1} \varepsilon_{t-\tau+1}$, obtaining the residuals $\hat{\varepsilon}_t$.
- 3 We simulate i.i.d variables η_t and calculate $\hat{v}_t^* = \eta_t \hat{\varepsilon}_t + \hat{\theta}_1 \eta_{t-1} \hat{\varepsilon}_{t-1} + \dots + \hat{\theta}_{\tau-1} \eta_{t-\tau+1} \hat{\varepsilon}_{t-\tau+1}$.
- 4 We estimate the parameter β_{k_0} of the model with $x_{k_0,t}$ and build the bootstrapped data $y_t^* = \hat{\beta}_{k_0} x_{k_0,t} + \hat{v}_t^*$.
- 5 With the sample $y_1^*, \dots, y_{T+P-\tau}^*$, calculate the statistics MSE-t-mM and MSE-t-Mm.

In point 3 we depart from Clark and McCracken (2012) 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. We took the idea from 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 $\{\text{MSE-t}_{ij}\}_{(i,j) \in I_0 \times J_0}$ and its bootstrapped counterpart $\{\text{MSE-t}_{ij}^*\}_{(i,j) \in I_0 \times J_0}$ have the same asymptotic distribution. 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 the mathematical appendix we show that this effect is asymptotically negligible. Here we only outline the argument: if $i \in I_0$ and $j \in J_0$, then $k_1 \subseteq i, j$ and thus, $k_1 \subseteq i, i \cap j \subseteq k_0$. 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 .

If the null does not hold, then k_1 is not included in k_0 . Consequently, by assumption 2, there is a model $j \in J$ such that $\forall i \in I, \sigma_\tau^2(j) < \sigma_\tau^2(i)$. Then, MSE-t-mM and MSE-t-Mm diverge to $+\infty$. On the other hand, let $k_2 = k_1 \cap k_0$. Then, the bootstrapped statistics will be asymptotically distributed as their counterparts under the null that $y_{t+\tau} = x'_{k_2, t} \beta_{k_2} + u_{k_2, t+\tau}$, so the critical values obtained by the bootstrap are bounded and the asymptotic power of the test is 1.

4.1 An extension

Suppose the object of interest is not just whether any model in J beats the benchmarks, but precisely which ones in J do that. If we express the null as a composite hypothesis $\cap_{j \in J} H_{0,j} = \cap_{j \in J} \{\sigma_\tau^2(j) \geq \min_{i \in I} \sigma_\tau^2(i)\}$, then the question is which $H_{0,j}$ are false. In the case of one benchmark, this can be determined by the procedure of Romano and Wolf (2005), while controlling the familywise error rate (FWE), that is, the probability that one true $H_{0,j}$ is rejected (the RC only guarantees that when all of them are true).

We can adapt their method as follows. First calculate

$$\hat{d}_1 = \inf\{x : P^*[\text{MSE-t-Mm}^* \leq x] \geq 1 - \alpha\},$$

where P^* is the bootstrap probability distribution. Then, we reject $H_{0,j}$ when

$$\sigma_\tau^2(j) < \min_{i \in I} \{\sigma_\tau^2(i) - \hat{d}_1(P - \tau + 1)^{1/2} \hat{S}_{d_{ij}}\}.$$

If $H_{0,j}$ is rejected for at least one j , then repeat the procedure excluding the indexes j of the already rejected hypotheses and continue until no hypothesis is rejected. The fact that the FWE is controlled can be proved by modifying the proof of theorem 3.1(ii) in Romano and Wolf (2005).

5 Monte Carlo

We have designed a Monte Carlo experiment with three different Data Generation Processes.

- DGP 1: this scenario is similar to DPG 1 of Clark and McCracken (2012), but including several univariate benchmarks with lags of y_t .
- DGP 2: similar to the DGP reported in section 4.1 of Hubrich and West (2010), but again, with benchmarks with lags of y_t .
- DGP 3: in this scenario we test causality of a predictand y_t by another variable x_t when $(x_t, y_t)'$ is generated according to a VAR model.

We report the results in two forms. First, in tables of rejection frequencies and second by the empirical cumulative distribution functions (ecdf) of the p-values. We run 2,000 realizations of the DGP and obtain as many p-values

with 500 bootstrap samples. Under the null hypothesis, the p-values should be distributed according to a uniform distribution in $(0, 1)$, so the ecdf would be near the diagonal. We represent the ecdfs only in the interval $[0, 0.2]$, since theoretical significance levels below 0.8 are not usual. Under an alternative hypothesis, the more powerful the test is, the further its ecdf is from the diagonal towards the top left.

In order to assess the power of the MSE-t-mM and MSE-t-Mm tests, we compare it to the results with the following procedure: we pick the benchmark i^* whose out-of-sample forecasts have the minimum MSE and then, we apply the single-benchmark RC test with the statistic $\max_j \text{MSE-t}_{i^*j}$.

5.1 DGP 1

This scenario is similar to DGP 1 in Clark and McCracken (2012), that is intended to replicate the real case of forecasting the US core inflation. We simulate the predictand y_t and a set of predictors $x_{m,t}$, with $m = 1, \dots, n$, with $n = 5$ (rather than $n = 7$ to reduce the computational burden) as

$$\begin{aligned} y_{t+\tau} &= -0.3y_t + bx_{1,t} + u_{t+\tau} \\ u_{t+\tau} &= \begin{cases} \varepsilon_t & \text{if } \tau = 1 \\ \varepsilon_t + 0.95\varepsilon_{t-1} + 0.9\varepsilon_{t-2} + 0.8\varepsilon_{t-3} & \text{if } \tau = 4 \end{cases} \\ z_{m,t} &= \gamma_i z_{m,t-1} + v_{m,t}, \gamma_m = 0.8 - 0.15(m-1), \end{aligned}$$

where $\mathbb{E}\varepsilon_t v_{m,t} = \mathbb{E}v_{m,t} v_{l,t} = 0$, $\text{Var}(\varepsilon_t) = 2$ and $\text{Var}(v_{m,t}) = 1 - \gamma_i^2$.

The main difference of our experiment is that we consider several univariate benchmarks instead of only one. Let $x_t = (y_t, \dots, y_{t-4}, x_{1,t}, \dots, x_{5,t})'$. We choose I as the collection of all sets $i = \{1, \dots, p\}$, with $p = 1, \dots, 4$, that is, all the autoregressive models with maximum lag p . On the other hand, J comprises all sets of the form $j = i \cup k$, where $k \subset \{5, \dots, 9\}$ and $k \neq \emptyset$, that is, we have $4(2^5 - 1) = 124$ bivariate models. The simulation is repeated for $P = R = 40, 80$ and 120. In tables 2 and 3 we report the rejection frequencies for $\alpha = 0.1$ and $\alpha = 0.05$ and in figure 1 we represent the ecdfs of the p-values just for the case $P = R = 80$.

5.2 DGP 2

The second scenario is based in the one in section 4.1 of Hubrich and West (2010) and is related to the question whether disaggregate data can be useful to predict an aggregate. Here it is assumed that the predictand y_t is the aggregation of three components $y_{i,t}$, $i = 1, 2, 3$. Here the benchmark is an univariate model of y_t and the alternative ones are bivariate models that include some component $y_{i,t}$. Unlike Hubrich and West, we do not focus on the case of small sets of models. Consequently, we assume that the true orders of the autoregressive models are unknown and we allow them to range from one lag to four. Thus, the benchmarks are

$$\hat{y}_{t+1} = c + \sum_{j=0}^p \beta_{0j} y_{t-j}$$

with $p = 0, \dots, 3$ and the alternatives are

$$\hat{y}_{t+1} = c + \sum_{j=0}^p \beta_{0j} y_{t-j} + \sum_{j=0}^q \beta_{1j} y_{i,t-j}$$

with $p, q = 0, \dots, 3$ and $i = 1, 2$. Hence, we have 4 benchmarks and 24 alternatives.

The simulations under the null use the same univariate autoregressive model for all three components $y_{i,t} = 1 + 0.5y_{i,t-1} + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ is standard Gaussian white noise. Under the alternative hypothesis, the vector $(y_{1,t}, y_{2,t}, y_{3,t})'$ is generated by a VAR(1) with the matrix

$$\begin{pmatrix} 0.5 & -0.6 & 0 \\ -0.4 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}.$$

The rejection frequencies under the null and the alternative are in tables 4 and 5. The cdfs of the p-values are in figure 2.

5.3 DGP 3

Now we want to check the performance of the test when the object of interest is a causality relation between two variables. For this, we generate data according

to the following model

$$\begin{aligned}y_{t+1} &= \phi_{11}y_t + \phi_{12}y_{t-1} + \phi_{13}y_{t-2} + bx_t + \xi_{1t} \\x_{t+1} &= \phi_{21}x_t + \phi_{22}x_{t-1} + \phi_{23}x_{t-2} + \xi_{2t},\end{aligned}$$

where $\phi_{1j} = 0.6, -0.3, 0.2$ and $\phi_{2j} = 0.4, 0.2, 0.1$. As in the previous cases, when $b = 0$, the null hypothesis holds. We report results for $b = 0$ and $b = 0.4$.

In this example, all the regressors are either lags of y_t or lags of x_t . The first class of models, I comprises the univariate autoregressive models that include p lags of y_t up to $p = 6$, whereas J comprises all the bivariate models with p lags of y_t and q lags of x_t , where p and q range from 1 to 6, so J has 36 models.

Otherwise, the conditions of the experiment are as in the previous cases. The rejection frequencies are in tables 6 and 7 and the cdfs of the p-values for $P = 80$ are in figure 3.

5.4 Discussion

The simulations show that under the null, the empirical size of both tests is acceptably close to the theoretical one. All the simulations show that MSE-t-Mm has greater power than MSE-t-mM and the difference is greater for long series. This is probably a consequence of the fact that for any matrix A , $\max_i \min_j A_{ij} \leq \min_j \max_i A_{ij}$. This implies that $\text{MSE-t-Mm} \leq \text{MSE-t-mM}$ and then, the critical values of MSE-t-Mm are lower than the ones of MSE-t-mM. On the other hand, both statistics appear to remain close when the null does not hold. In simulations, their cdfs under the alternative hypothesis are closer than under the null and in fact, it can be proved that when the null does not hold, $\text{MSE-t-mM} = \text{MSE-t-Mm} + O_p(1)$.

As expected, the power of the test increases with the length of the series, which is consistent with the theoretical result that asymptotically, the power of both tests is 1. Again, as expected, even MSE-t-mM is more powerful than $\max_j \text{MSE-t-}i^*j$.

6 Real data example

In this section, we apply our tests to the problem of determining whether the unemployment is useful to predict inflation one step ahead. The series used in this application are the Unemployment Rate (UNEM) and the Consumer Price Index (CPI) of USA. Both series are subject to preliminary transformations, so $y_t = \nabla \nabla \log \text{CPI}_t$ and $x_t = \nabla \log \text{UNEM}_t$, where $\nabla u_t = u_t - u_{t-1}$. The length of bivariate series, after the transformation is $T = 648$ (from 1958:1 to 2010:12).

Since both series have some seasonal component, it is convenient to use seasonal models. Thus, our univariate benchmarks are

$$\hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} \quad \hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} + \sum_{j=s-1}^{s+p} \beta_{0,j} y_{t-j}, \quad (5)$$

where $s = 12$ and p ranges from 0 to 6. The alternative models are

$$\hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} + \sum_{j=0}^q \beta_{1,j} x_{t-j} \quad (6)$$

$$\hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} + \sum_{j=s-1}^{p+s} \beta_{0,j} y_{t-j} + \sum_{j=0}^q \beta_{1,j} x_{t-j} \quad (7)$$

$$\hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} + \sum_{j=0}^q \beta_{1,j} x_{t-j} + \sum_{j=s-1}^{s+q} \beta_{1,j} x_{t-j} \quad (8)$$

$$\hat{y}_{t+\tau} = \sum_{j=0}^p \beta_{0,j} y_{t-j} + \sum_{j=s-1}^{p+s} \beta_{0,j} y_{t-j} + \sum_{j=0}^q \beta_{1,j} x_{t-j} + \sum_{j=s-1}^{s+q} \beta_{1,j} x_{t-j} \quad (9)$$

where p and q range from 0 to 6.

We have computed the statistics MSE-t-mM, MSE-t-Mm and $\max_j \text{MSE-t}_{i^*j}$, with $M = 5000$ replications for the bootstrap and $\lambda = 0.5$. The prediction errors of the full model present some skewness, so in the bootstrap we use the distribution with $\mathbb{E}\eta^3 = 1$. The p-values are, 0.04 for MSE-t-Mm, 0.07 for MSE-t-mM .

Let us see now some of the risks of the RC with a unique benchmark. In table 1 we represent for each univariate model of (5), the p-value of the single-benchmark Reality Check against all the alternatives in a set that comprises those models in (6)–(9) that include exactly the same lags of y_t as the benchmark.

Below, we reproduce the MSEs of the benchmark and the best bivariate model. What we get is that the result of the test is strongly dependent on the particular benchmark. Sometimes, we reject the null strongly, but we can see that this happens because our benchmark is not very good. We may prevent this by picking the benchmark that has smallest MSE, but this is the naïve method that we tried in the Monte Carlo section. Then, we saw that it is less powerful than our tests and this is reflected in its 0.15 p-value.

Table 1: p-values of the single-benchmark RC.

AR order	1	2	3	4	5
p-value	0	0	0.1556	0.0928	0.1862
MSE_i	0.8525	0.7949	0.6520	0.5168	0.4715
$\min_{j \in J} MSE_j$	0.7730	0.7256	0.6132	0.4816	0.4442
AR order (seasonal)	1	2	3	4	5
p-value	0	0.0822	0.3976	0.0732	0.1526
MSE_i	0.7855	0.6805	0.5338	0.4676	0.4402
$\min_{j \in J} MSE_j$	0.7330	0.6558	0.5080	0.4332	0.4120

7 Conclusions

There are situations in which we want to compare a class of models to a benchmark from a given class, but it is not clear which one should be chosen. Rather than to pick up one, it is preferable to take all of them into account. To do this, we use a straightforward generalization of the statistic of the RC for nested models and a modification of the semiparametric wild bootstrap. The novelty of our bootstrap is that it generates the artificial samples assuming that the true model is one that nests all the possible benchmarks, while satisfying the null hypothesis.

We show with Monte Carlo simulations in different scenarios that our method performs better than just the RC with the better-performing benchmark. We also apply our test to show that there is considerable evidence that bivariate models that use unemployment can predict better the inflation than univariate

ones.

A Mathematical Appendix

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 $h_t = x_t u_t$, $H(t) = t^{-1} \sum_{s=1}^{t-\tau} h_t$ and \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 $\tilde{h}_{ij,t} = \sigma^{-1} \tilde{A}_{ij} B^{1/2} h_t$. 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, $S_{ij} = \sigma^{-2} S_{hh}^{1/2} B^{1/2} \tilde{A}'_{ij} \tilde{A}_{ij} B^{1/2} S_{hh}^{1/2}$.

The technical assumptions that were not discussed in section 2 are the following.

Assumption 3. (a) $U_{t+\tau} = [h'_{t+\tau}, \text{vec}(x_t x_t' - \mathbb{E}x_t x_t')]'$ 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) $\lim_T T^{-1} \mathbb{E}(\sum_{s=1}^{T-\tau} U_{s+\tau})(\sum_{s=1}^{T-\tau} U_{s+\tau})' = \Omega < \infty$ 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. $\lim_{P,T} P/T = \lambda_P \in (0, \infty)$.

A.1 Asymptotics of MSE-t-Mm and MSE-t-mM

To determine the asymptotic distributions under the null, 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 1. $\text{MSE-t-mM} = \min_{i \in I_0} \max_{j \in J_0} \text{MSE-t}_{ij} + o_p(1)$ and $\text{MSE-t-Mm} = \max_{j \in J_0} \min_{i \in I_0} \text{MSE-t}_{ij} + o_p(1)$.

Proof. We will prove the result for the MSE-t-mM and leave the other case to the reader. First

$$\text{MSE-t-mM} = \min \left(\min_{i \in I_0} \max_{j \in J} \text{MSE-t}_{ij}, \min_{i \in I \setminus I_0} \max_{j \in J} \text{MSE-t}_{ij} \right).$$

Since $\forall i \in I \setminus I_0, \max_{j \in J} \text{MSE-t}_{ij} \xrightarrow{P} \infty$, then $\min_{i \in I \setminus I_0} \max_{j \in J} \text{MSE-t}_{ij} \xrightarrow{P} \infty$. Thus, by lemma 1,

$$\text{MSE-t-mM} = \min_{i \in I_0} \max_{j \in J} \text{MSE-t}_{ij} + o_p(1). \quad (10)$$

Now, we can apply again lemma 1 for each i to get $\max_{j \in J} \text{MSE-t}_{ij} = \max_{j \in J_0} \text{MSE-t}_{ij} + o_p(1)$. Then, we conclude by replacing in (10) and invoking the continuous mapping theorem. \square

Proposition 2. *The asymptotic distributions of the tests are given by*

$$\text{MSE-t-mM} \xrightarrow{d} \min_{i \in I_0} \max_{j \in J_0} g_{ij}, \quad (11)$$

$$\text{MSE-t-Mm} \xrightarrow{d} \max_{j \in J_0} \min_{i \in I_0} g_{ij}, \quad (12)$$

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

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

and $W(\omega)$ is a $n \times 1$ standard Brownian motion.

Proof. 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 (2011) when i and j are overlapping. We only need to be careful to ensure that the Brownian motion 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

$$\begin{aligned} \sum_{s=[ut]}^t \tilde{H}(s)\tilde{h}'_s &= \sigma^{-2}\tilde{A}B^{1/2}C \sum_s^t K(s)k'_s C' B^{1/2}' \tilde{A}' \Rightarrow \\ &\sigma^{-2}\tilde{A}B^{1/2}C \left\{ \int_u^1 W(\omega)dW(\omega)' \right\} C' B^{1/2}' \tilde{A}'. \end{aligned}$$

Consequently,

$$\begin{aligned} \sum_{s=[ut]}^t \tilde{h}'_s \tilde{H}(s) &= \text{tr} \sum_{s=[ut]}^t \tilde{H}(s)\tilde{h}'_s \Rightarrow \\ \sigma^{-2}\text{tr}\tilde{A}B^{1/2}C \left\{ \int_u^1 W(\omega)dW(\omega)' \right\} C' B^{1/2}' \tilde{A}' &= \\ \int_u^1 W(\omega)' \left(\sigma^{-2}C' B^{1/2}' \tilde{A}' \tilde{A}B^{1/2}C \right) dW(\omega). \end{aligned}$$

□

As stated in the discussion of the Monte Carlo, the asymptotic behavior of both tests under the alternative hypothesis satisfies the constraint

$$\text{MSE-t-mM} = \text{MSE-t-Mm} + O_p(1).$$

This is a consequence of the fact that when $\sigma_\tau^2(i) < \sigma_\tau^2(j)$, then $(P - \tau + 1)^{1/2}\text{MSE-t}_{ij} \xrightarrow{p} (\sigma_\tau^2(i) - \sigma_\tau^2(j))/S_{d_{ij}}$, $S_{d_{ij}}$ is the population counterpart of $\hat{S}_{d_{ij}}$. This entails that

$$(P - \tau + 1)^{-1/2}\text{MSE-t-mM} \xrightarrow{p} \min_{i \in I_0} \max_{j \in J_0} (\sigma_\tau^2(i) - \sigma_\tau^2(j))/S_{d_{ij}} \quad (13)$$

$$(P - \tau + 1)^{-1/2}\text{MSE-t-Mm} \xrightarrow{p} \max_{j \in J_0} \min_{i \in I_0} (\sigma_\tau^2(i) - \sigma_\tau^2(j))/S_{d_{ij}}. \quad (14)$$

In fact, since $\hat{u}_{j,t}$ approach to the unique u_t as $\hat{\beta}_{j,t} \rightarrow \beta_j^*$, $S_{d_{ij}}$ only depends on i . On the other hand, for any $a_i, b_j, c_i \in \mathbb{R}$, $\min_i \max_j (a_i - b_j)/c_i = \max_j \min_i (a_i - b_j)/c_i$. Hence, the constants in the right hand sides of (13) and (14) are the same. Since $\hat{d}_{i,j} = \sigma_\tau^2(i) - \sigma_\tau^2(j) + O_p(T^{-1/2})$ and $\hat{S}_{d_{ij}} = S_{d_{ij}} + O_p(T^{-1/2})$, then

$$\text{MSE-t-mM} - \text{MSE-t-Mm} = (P - \tau + 1)^{1/2}O_p(T^{-1/2}) = O_p(1).$$

A.2 Validity of the bootstrap

We have to prove that $\text{MSE-t}_{ij} \xrightarrow{d^*} g_{ij}^*$, where $g_{ij}^* \sim (\Gamma_{1,ij}^* - 0.5\Gamma_{2,ij}^*)/\Gamma_{3,ij}^{*1/2}$, where $\Gamma_{m,ij}^*$ is distributed as $\Gamma_{m,ij}$, for $m = 1, 2, 3$. Let us compare our bootstrap with those of Clark and McCracken (2011) and (2012).

- 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 $k_1 \subset i \subset k_0$, 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 since $k_1 \subset i \subset k_0$, then $k_1 \subset i \cap j \subset k_0$ and again, the excess parameters have null population values.

We have to see that the excess regressors in the artificial sample only apert a $o_p^*(1)$ error to the MSE-t_{ij}^* statistics. Let us check this for the numerator of MSE-t_{ij}^* . For the denominator, 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

$$\sum_s \left\{ (\hat{u}_{i,s+\tau}^*(i))^2 - \hat{u}_{j,s+\tau}^*(i)^2 \right\} - (\hat{u}_{i,s+\tau}^*(k_0))^2 - \hat{u}_{j,s+\tau}^*(k_0))^2 \Big\} = o_p(1). \quad (15)$$

In order to see that (15) 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,

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

Now, that

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

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

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

and

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

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

B Tables and figures

Table 2: DGP 1: rejection frequencies under the null ($b = 0$).

	P=40		P=80		P=120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
	$\tau = 1$					
MSE-t-mM	0.0795	0.0410	0.0895	0.0475	0.1085	0.0515
MSE-t-Mm	0.0850	0.0475	0.0975	0.0600	0.1095	0.0570
MSE-t-naïve	0.0645	0.0300	0.0625	0.0300	0.0815	0.0410
	$\tau = 4$					
MSE-t-mM	0.1120	0.0685	0.1075	0.0560	0.1250	0.0695
MSE-t-Mm	0.1250	0.0735	0.1085	0.0615	0.1240	0.0700
MSE-t-naïve	0.0925	0.0530	0.0770	0.0390	0.0945	0.0525

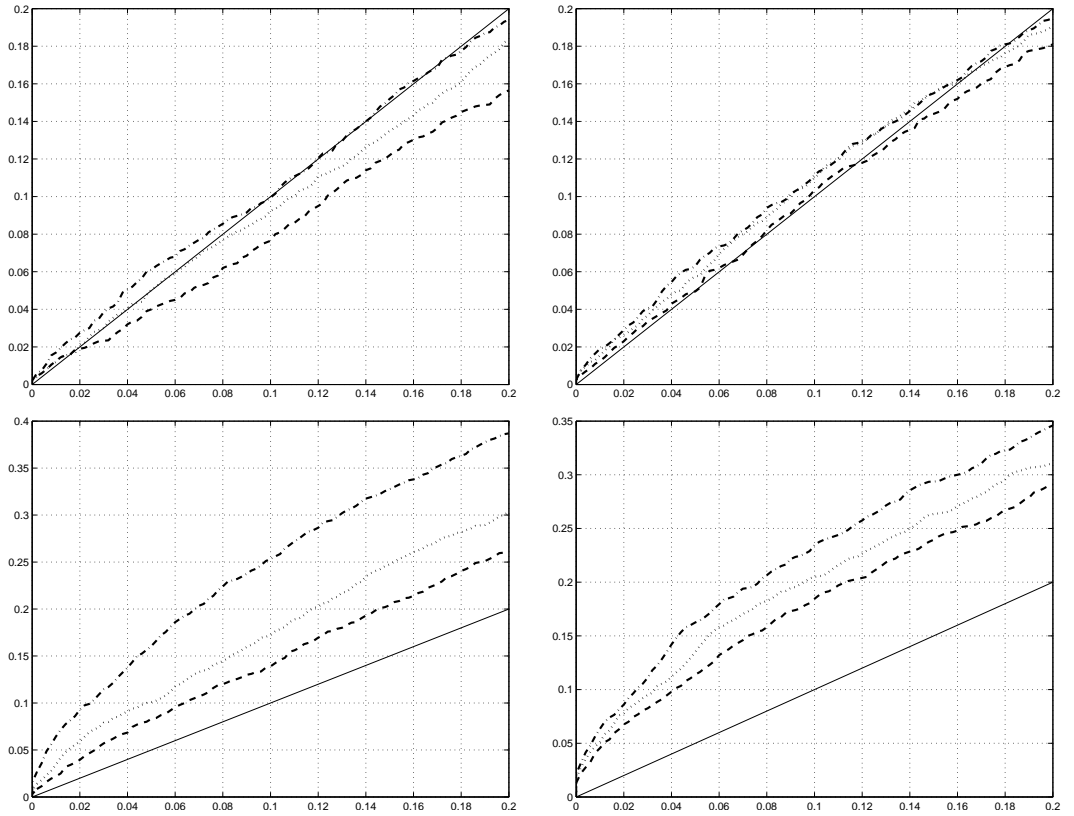


Figure 1: DGP 1: results under the null hypothesis (above) and the alternative (below), for $P = R = 80$. Left is $\tau = 1$ and the right is $\tau = 4$. MSE-t-mM is represented by the dotted line, MSE-t-Mm by the dash-dot line and the naïve test by the dashed line.

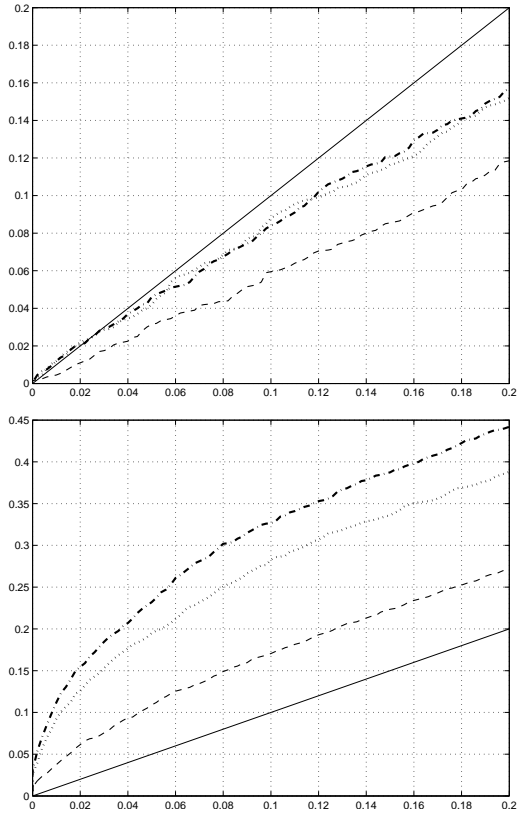


Figure 2: DGP 2: results under the null hypothesis (above) and the alternative (below), for $P = R = 80$. MSE-t-mM is represented by the dotted line, MSE-t-Mm by the dash-dot line and the naïve test by the dashed line.

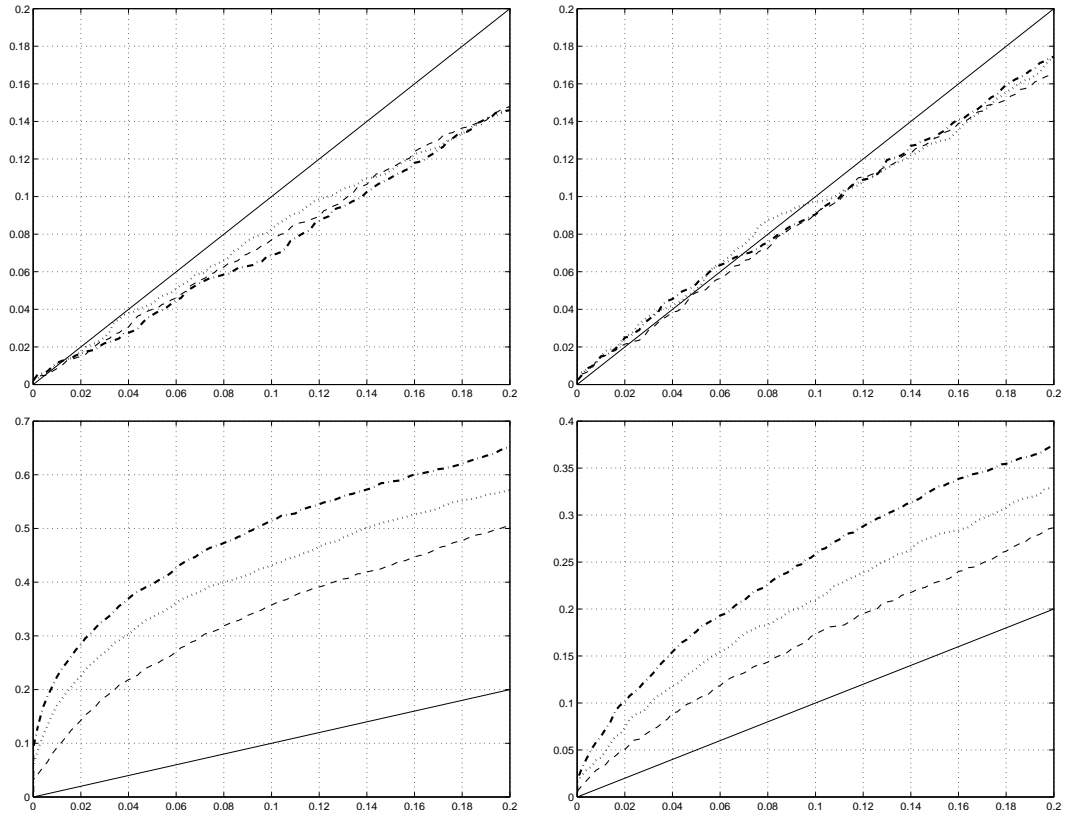


Figure 3: DGP 3: results under the null hypothesis (above) and the alternative (below), for $P = R = 80$. Left is $\tau = 1$ and right, $\tau = 3$. MSE-t-mM is represented by the dotted line, MSE-t-Mm by the dash-dot line and the naïve test by the dashed line.

References

- [1] Clark T. E. and McCracken M. W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.
- [2] Clark T. E. and McCracken M. W. (2005), "Evaluating Direct Multistep

Table 3: DGP 1: rejection frequencies under the alternative ($b = 0.4, 0.8$).

	P=40		P=80		P=120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
$\tau = 1$						
MSE-t-mM	0.1155	0.0745	0.1705	0.1000	0.2430	0.1710
MSE-t-Mm	0.1530	0.0910	0.2510	0.1565	0.3435	0.2320
MSE-t-naïve	0.0825	0.0380	0.1110	0.0650	0.1695	0.1015
$\tau = 4$						
MSE-t-mM	0.1570	0.0890	0.2035	0.1295	0.2695	0.1805
MSE-t-Mm	0.1680	0.1080	0.2290	0.1600	0.2890	0.1990
MSE-t-naïve	0.1030	0.0575	0.1515	0.0895	0.2045	0.1265

Table 4: DGP 2: rejection frequencies under the null.

	P =40		P =80		P =120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
MSE-t-mM	0.0515	0.0285	0.0850	0.0400	0.0905	0.0490
MSE-t-Mm	0.0585	0.0290	0.0825	0.0425	0.1000	0.0505
MSE-t-naïve	0.0405	0.0185	0.0590	0.0300	0.0675	0.0345

Table 5: DGP 2: rejection frequencies under the alternative.

	P =40		P =80		P =120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
MSE-t-mM	0.1290	0.0740	0.2765	0.1900	0.3910	0.2875
MSE-t-Mm	0.1600	0.0930	0.3260	0.2280	0.4485	0.3470
MSE-t-naïve	0.0775	0.0390	0.1685	0.1050	0.2435	0.1650

Forecasts”, *Econometric Reviews* 24, 369-404.

[3] Clark T. E. and McCracken M. W. (2011), ”Tests of Equal Forecast Accuracy for Overlapping Models”. Federal Reserve Bank of St. Louis, Working Paper Series.

[4] Clark T. E. and McCracken M. W. (2012), ”Reality Checks and Comparisons of Nested Predictive Models,” *Journal of Business and Economic Statistics* 30, 53-66.

Table 6: DGP 3: rejection frequencies under the null ($b = 0$).

	P=40		P=80		P=120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
$\tau = 1$						
MSE-t-mM	0.0670	0.0345	0.0805	0.0420	0.0845	0.0410
MSE-t-Mm	0.0670	0.0365	0.0680	0.0355	0.0800	0.0375
MSE-t-naïve	0.0495	0.0245	0.0530	0.0280	0.0550	0.0235
$\tau = 3$						
MSE-t-mM	0.0865	0.0525	0.0970	0.0485	0.0925	0.0525
MSE-t-Mm	0.0900	0.0575	0.0895	0.0515	0.1015	0.0530
MSE-t-naïve	0.0645	0.0330	0.0645	0.0325	0.0660	0.0325

Table 7: DGP 3: rejection frequencies under the alternative ($b = 0.4$).

	P=40		P=80		P=120	
	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.9$	$\alpha = 0.95$
$\tau = 1$						
MSE-t-mM	0.1945	0.1350	0.4280	0.3315	0.6415	0.5195
MSE-t-Mm	0.2315	0.1670	0.5105	0.3925	0.7145	0.6075
MSE-t-naïve	0.1260	0.0790	0.2975	0.1955	0.4840	0.3690
$\tau = 3$						
MSE-t-mM	0.1160	0.0710	0.2080	0.1330	0.2655	0.1695
MSE-t-Mm	0.1355	0.0870	0.2545	0.1725	0.3260	0.2265
MSE-t-naïve	0.0775	0.0440	0.1355	0.0760	0.1695	0.1030

- [5] Clark T. E. and West K. D. (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.
- [6] Davidson, R. and Flachaire, E. (2008) "The wild bootstrap tamed at last" *Journal of Econometrics* 146, 162–169.
- [7] Diebold F. and Mariano R. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 252-263.
- [8] Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23, 365-380.
- [9] Hansen, B. E. (1992) "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes", *Econometric Theory* 8, 489-500.
- [10] Hubrich, K., and West, K. D. (2010) ."Forecast Evaluation of Small Nested Model Sets", *Journal of Applied Econometrics* 25, 574-594.
- [11] McCracken, M. W. (2004), "Asymptotics for OutofSample Tests of Causality," manuscript, University of Missouri.
- [12] Pincheira, P. (2011), "A bunch of models, a bunch of nulls and inference about predictive ability", Working Paper 607, Central Bank of Chile.
- [13] Politis, D. N. and Romano, J.P. (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association* 89, 1303-1313.
- [14] Romano, J. P. and Wolf, M. (2005) "Stepwise multiple testing as formalized data snooping," *Econometrica* 73, 1237–1282.
- [15] Vuong, Q. H. (1989). "Likelihood ratio tests for model selection and non-nested hypotheses". *Econometrica* 57, 307-333.
- [16] West, K. D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica* 64, 1067-1084.
- [17] White, H. (2000), "A Reality Check for Data Snooping," *Econometrica* 68, 1098–1126.