Out-of-Sample Forecast Tests Robust to the Choice of Window Size

Barbara Rossi and Atsushi Inoue

(ICREA, UPF, CREI, BGSE, Duke) (NC State)

April 1, 2012

Abstract

This paper proposes new methodologies for evaluating out-of-sample forecasting performance that are robust to the choice of the estimation window size. The methodologies involve evaluating the predictive ability of forecasting models over a wide range of window sizes. We show that the tests proposed in the literature may lack the power to detect predictive ability and might be subject to data snooping across different window sizes if used repeatedly. An empirical application shows the usefulness of the methodologies for evaluating exchange rate models' forecasting ability.

Keywords: Predictive Ability Testing, Forecast Evaluation, Estimation Window.

Acknowledgments: We thank the editor, the associate editor, two referees as well as S. Burke, M.W. McCracken, J. Nason, A. Patton, K. Sill, D. Thornton and seminar participants at the 2010 Econometrics Workshop at the St. Louis Fed, Bocconi University, U. of Arizona, Pompeu Fabra U., Michigan State U., the 2010 Triangle Econometrics Conference, the 2011 SNDE Conference, the 2011 Conference in honor of Hal White, the 2011 NBER Summer Institute and the 2011 Joint Statistical Meetings for useful comments and suggestions. This research was supported by National Science Foundation grants SES-1022125 and SES-1022159 and North Carolina Agricultural Research Service Project NC02265.

J.E.L. Codes: C22, C52, C53

1 Introduction

This paper proposes new methodologies for evaluating the out-of-sample forecasting performance of economic models. The novelty of the methodologies that we propose is that they are robust to the choice of the estimation and evaluation window size. The choice of the estimation window size has always been a concern for practitioners, since the use of different window sizes may lead to different empirical results in practice. In addition, arbitrary choices of window sizes have consequences about how the sample is split into in-sample and out-of-sample portions. Notwithstanding the importance of the problem, no satisfactory solution has been proposed so far, and in the forecasting literature it is common to only report empirical results for one window size. For example, to illustrate the differences in the window sizes, we draw on the literature on forecasting exchange rates (the empirical application we will focus on): Meese and Rogoff (1983a) use a window of 93 observations in monthly data, Chinn (1991) a window size equal to 45 in quarterly data, Qi and Wu (2003) use a window of 216 observations in monthly data, Cheung et al. (2005) consider windows of 42 and 59 observations in quarterly data, Clark and West's (2007) window is 120 observations in monthly data, Gourinchas and Rey (2007) consider a window of 104 observations in quarterly data, and Molodtsova and Papell (2009) consider a window size of 120 observations in monthly data. This common practice raises two concerns. A first concern is that the "ad hoc" window size used by the researcher may not detect significant predictive ability even if there would be significant predictive ability for some other window size choices. A second concern is the possibility that satisfactory results were obtained simply by chance, after data snooping over window sizes. That is, the successful evidence in favor of predictive ability might have been found after trying many window sizes, although only the results for the successful window size were reported and the search process was not taken into account when evaluating their statistical significance. Only rarely do researchers check the robustness of the empirical results to the choice of the window size by reporting results for a selected choice of window sizes. Ultimately, however, the size of the estimation window is not a parameter of interest for the researcher: the objective is rather to test predictive ability and, ideally, researchers would like to reach empirical conclusions that are robust to the choice of the estimation window size.

This paper views the estimation window as a "nuisance parameter": we are not interested in selecting the "best" window; rather we would like to propose predictive ability tests that are "robust" to the choice of the estimation window size. The procedures that we

propose ensure that this is the case by evaluating the models' forecasting performance for a variety of estimation window sizes, and then taking summary statistics of this sequence. Our methodology can be applied to most tests of predictive ability that have been proposed in the literature, such as Diebold and Mariano (1995), West (1996), McCracken (2000) and Clark and McCracken (2001). We also propose methodologies that can be applied to Mincer and Zarnowitz's (1969) tests of forecast efficiency, as well as more general tests of forecast optimality. Our methodologies allow both for rolling as well as recursive window estimation schemes and let the window size to be large relative to the total sample size. Finally, we also discuss methodologies that can be used in the Giacomini and White's (2005) and Clark and West's (2007) frameworks, where the estimation scheme is based on a rolling window with fixed size.

This paper is closely related to the works by Pesaran and Timmermann (2007) and Clark and McCracken (2009), and more distantly related to Pesaran, Pettenuzzo and Timmermann (2006) and Giacomini and Rossi (2010). Pesaran and Timmermann (2007) propose cross validation and forecast combination methods that identify the "ideal" window size using sample information. In other words, Pesaran and Timmermann (2007) extend forecast averaging procedures to deal with the uncertainty over the size of the estimation window, for example, by averaging forecasts computed from the same model but over various estimation window sizes. Their main objective is to improve the model's forecast. Similarly, Clark and McCracken (2009) combine rolling and recursive forecasts in the attempt to improve the forecasting model. Our paper instead proposes to take summary statistics of tests of predictive ability computed over several estimation window sizes. Our objective is not to improve the forecasting model nor to estimate the ideal window size. Rather, our objective is to assess the robustness of conclusions of predictive ability tests to the choice of the estimation window size. Pesaran, Pettenuzzo and Timmermann (2006) have exploited the existence of multiple breaks to improve forecasting ability; in order to do so, they need to estimate the process driving the instability in the data. An attractive feature of the procedure we propose is that it does not need to impose nor determine when the structural breaks have happened. Giacomini and Rossi (2010) propose techniques to evaluate the relative performance of competing forecasting models in unstable environments, assuming a "given" estimation window size. In this paper, our goal is instead to ensure that forecasting ability tests be robust to the choice of the estimation window size. That is, the procedures that we propose in this paper are designed for determining whether findings of predictive ability are robust to the choice of the window size, not to determine which point in time the predictive ability shows up:

the latter is a very different issue, important as well, and was discussed in Giacomini and Rossi (2010). Finally, this paper is linked to the literature on data snooping: if researchers report empirical results for just one window size (or a couple of them) when they actually considered many possible window sizes prior to reporting their results, their inference will be incorrect. This paper provides a way to account for data snooping over several window sizes and removes the arbitrary decision of the choice of the window length.

After the first version of this paper was submitted, we became aware of independent work by Hansen and Timmermann (2011). Hansen and Timmermann (2011) propose a sup-type test similar to ours, although they focus on p-values of the Diebold and Mariano's (1995) test statistic estimated via a recursive window estimation procedure for nested models' comparisons. They provide analytic power calculations for the test statistic. Our approach is more generally applicable: it can be used for inference on out-of-sample models' forecast comparisons and to test forecast optimality where the estimation scheme can be either rolling, fixed or recursive, and the window size can be either a fixed fraction of the total sample size or finite. Also, Hansen and Timmermann (2011) do not consider the effects of time-varying predictive ability on the power of the test.

We show the usefulness of our methods in an empirical analysis. The analysis re-evaluates the predictive ability of models of exchange rate determination by verifying the robustness of the recent empirical evidence in favor of models of exchange rate determination (e.g., Molodtsova and Papell, 2009, and Engel, Mark and West, 2007) to the choice of the window size. Our results reveal that the forecast improvements found in the literature are much stronger when allowing for a search over several window sizes. As shown by Pesaran and Timmermann (2005), the choice of the window size depends on the nature of the possible model instability and the timing of the possible breaks. In particular, a large window is preferable if the data generating process is stationary but comes at the cost of lower power, since there are fewer observations in the evaluation window. Similarly, a shorter window may be more robust to structural breaks, although it may not provide as precise an estimation as larger windows if the data are stationary. The empirical evidence shows that instabilities are widespread for exchange rate models (see Rossi, 2006), which might justify why in several cases we find improvements in economic models' forecasting ability relative to the random walk for small window sizes.

The paper is organized as follows. Section 2 proposes a framework for tests of predictive ability when the window size is a fixed fraction of the total sample size. Section 3 presents tests of predictive ability when the window size is a fixed constant relative to the total sample

size. Section 4 shows some Monte Carlo evidence on the performance of our procedures in small samples, and Section 4 presents the empirical results. Section 5 concludes.

2 Robust Tests of Predictive Accuracy When the Window Size is Large

Let $h \geq 1$ denote the (finite) forecast horizon. We assume that the researcher is interested in evaluating the performance of $h-$steps-ahead direct forecasts for the scalar variable y_{t+h} using a vector of predictors x_t using either a rolling, recursive or fixed window direct forecast scheme. We assume that the researcher has P out-of-sample predictions available, where the first prediction is made based on an estimate from a sample $1, 2, ..., R$, such that the last out-of-sample prediction is made based on an estimate from a sample of $T-R+1, ..., R+P-1 = T$ where $R+P+h-1 = T+h$ is the size of the available sample. The methods proposed in this paper can be applied to out-of-sample tests of equal predictive ability, forecast rationality and unbiasedness.

In order to present the main idea underlying the methods proposed in this paper, let us focus on the case where researchers are interested in evaluating the forecasting performance of two competing models: Model 1, involving parameters θ, and Model 2, involving parameters γ. The parameters can be estimated either with a rolling, fixed or a recursive window estimation scheme. In the rolling window forecast method, the true but unknown model's parameters θ^* and γ^* are estimated by $\widehat{\theta}_{t,R}$ and $\widehat{\gamma}_{t,R}$ using samples of R observations dated $t-R+1, ..., t$, for $t = R, R+1, ..., T$. In the recursive window estimation method, the model's parameters are instead estimated using samples of t observations dated $1, ..., t$, for $t = R, R+1, ..., T$. In the fixed window estimation method, the model's parameters are estimated only once using observations dated $1, ..., R$. Let $\left\{L_{t+h}^{(1)}\left(\widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$ and $\left\{L_{t+h}^{(2)}\left(\widehat{\gamma}_{t,R}\right)\right\}_{t=R}^{T}$ denote the sequence of loss functions of models 1 and 2 evaluating $h-$steps-ahead relative out-of-sample forecast errors, and let $\left\{\Delta L_{t+h}\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right\}_{t=R}^{T}$ denote their difference.

Typically, researchers rely on the Diebold and Mariano (1995), West (1996), McCracken (2000) or Clark and McCracken's (2001) test statistics for inference on the forecast error loss differences. For example, in the case of the Diebold and Mariano's (1995) and West's (1996) test, researchers evaluate the two models using the sample average of the sequence of

standardized out-of-sample loss differences:

$$\Delta L_T(R) \equiv \frac{1}{\widehat{\sigma}_R} P^{-1/2} \sum_{t=R}^{T} \Delta L_{t+h}(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}), \tag{1}$$

where $\widehat{\sigma}_R^2$ is a consistent estimate of the long run variance matrix of the out-of-sample loss differences, which differs in the Diebold and Mariano's (1995) and West's (1996) approaches.

The problem we focus on is that inference based on eq. (1) relies crucially on R, which is the size of the rolling window in the rolling estimation scheme or the way the sample is split into the in-sample and out-of-sample portions in the fixed and recursive estimation schemes. In fact, any out-of-sample test for inference regarding predictive ability does require researchers to choose R. The problem we focus on is that it is possible that, in practice, the choice of R may affect the empirical results. Our main goal is to design procedures that will allow researchers to make inference about predictive ability in a way that does not depend on the choice of the window size.

We argue that the choice of R raises two types of concerns. First, if the researcher tries several window sizes and then reports the empirical evidence based on the window size that provides him the best empirical evidence in favor of predictive ability, his test may be over-sized. That is, the researcher will reject the null hypothesis of equal predictive ability in favor of the alternative that the proposed economic model forecasts the best too often, thus finding predictive ability even if it is not significant in the data. The problem is that the researcher is effectively "data-mining" over the choice of R, and does not correct the critical values of the test statistic to take into account the search over window sizes. This is mainly a size problem.

A second type of concern arises when the researchers has simply selected an ad-hoc value of R without trying alternative values. In this case, it is possible that, when there is some predictive ability only over a portion of the sample, he may lack to find empirical evidence in favor of predictive ability because the window size was either too small or large to capture it. This is mainly a lack of power problem.

Our objective is to consider R as a nuisance parameter, and develop test statistics to perform inference about predictive ability that does not depend on R. The main results in this paper follow from a very simple intuition: if partial sums of the test function (either forecast error losses, or adjusted forecast error losses, or functions of forecast errors) obey a Functional Central Limit Theorem (FCLT), we can take any summary statistic across window sizes to robustify inference and derive its asymptotic distribution by applying the

Continuous Mapping Theorem (CMT). We consider two appealing and intuitive types of weighting schemes over the window sizes. The first scheme is to choose the largest value of the $\Delta L_T(R)$ test sequence, which corresponds to a "sup-type" test. This mimics to the case of a researcher experimenting with a variety of window sizes and reporting only the empirical results corresponding to the best evidence in favor of predictive ability. The second scheme involves taking a weighted average of the $\Delta L_T(R)$ tests, giving equal weight to each test. This choice is appropriate when researchers have no prior information on which window sizes are the best for their analysis. This choice corresponds to an average-type test. Alternative choices of weighting functions could be entertained and the asymptotic distribution of the resulting test statistics could be obtained by arguments similar to those discussed in this paper.

The following proposition states the general intuition behind the approach proposed in this paper. In the subsequent sub-sections we will verify that the high-level assumption in Proposition 1, eq. (2), holds for the test statistics we are interested in.

Proposition 1 (Asymptotic Distribution.) *Let $S_T(R)$ denote a test statistic with window size R. We assume that the test statistic $S_T(\cdot)$ we focus on satisfies*

$$S_T([\iota(\cdot)T]) \;\Rightarrow\; S(\cdot) \tag{2}$$

where $\iota(\cdot)$ is the identity function, that is, $\iota(x) = x$, and $\Rightarrow$ denotes weak convergence in the space of cadlag functions on $[0,1]$ equipped with the Skorokhod metric. Then,

$$\sup_{[\underline{\mu}T] \leq R \leq [\overline{\mu}T]} S_T(R) \xrightarrow{d} \sup_{\underline{\mu} \leq \mu \leq \overline{\mu}} S(\mu), \tag{3}$$

$$\frac{1}{[\overline{\mu}T] - [\underline{\mu}T] + 1} \sum_{R=[\underline{\mu}T]}^{[\overline{\mu}T]} S_T(R) \xrightarrow{d} \int_{\underline{\mu}}^{\overline{\mu}} S(\mu) d\mu \tag{4}$$

where $0 < \underline{\mu} < \overline{\mu} < 1$.

Note that this approach assumes that R is growing with the sample size and, asymptotically, becomes a fixed fraction of the total sample size. This assumption is consistent with the approaches by West (1996), West and McCracken (1998) and McCracken (2001). The next section will consider test statistics where the window size is fixed. Note also that based on Proposition 1 we can construct both one-sided as well as two-sided test statistics; for example, as a corollary of the Proposition, one can construct two-sided test statistics in

the "sup-type" test statistic by noting that $\sup_{[\underline{\mu}T]\leq R\leq[\overline{\mu}T]}|S_T(R)| \xrightarrow{d} \sup_{\underline{\mu}\leq\mu\leq\overline{\mu}}|S(\mu)|$, and similarly of the average-type test statistic.

In the existing tests, $\mu = \lim_{T\to\infty}\frac{R}{T}$ is fixed and condition (2) holds pointwise for a given μ. Condition (2) requires that the convergence holds uniformly in μ rather than pointwise, however. It turns out that this high-level assumption can be shown to hold for many of the existing tests of interest under their original assumptions. As we will show in the next subsections, this is because existing tests had already imposed assumptions for the FCLT to take into account recursive, rolling and fixed estimation schemes and because weak convergence to stochastic integrals can hold for partial sums (Hansen, 1992).

Note also that the practical implementation of (3) and (4) requires researchers to choose $\underline{\mu}$ and $\overline{\mu}$. To avoid data snooping over the choices of $\underline{\mu}$ and $\overline{\mu}$, we recommend researchers to impose symmetry by fixing $\overline{\mu} = 1-\underline{\mu}$, and to use $\underline{\mu} = [0.15]$ in practice. The recommendation is based on the small sample performance of the test statistics that we propose, discussed in Section 4.

We next discuss how this result can be directly applied to widely used measures of relative forecasting performance, where the loss function is the difference of the forecast error losses of two competing models. We consider two separate cases, depending on whether the models are nested or non-nested. Subsequently we present results for regression-based tests of predictive ability, such as Mincer and Zarnowitz's (1969) forecast rationality regressions, among others. For each of the cases that we consider, Appendix A in Inoue and Rossi (2012) provides a sketch of the proof that the test statistics satisfy condition (2) provided the variance estimator converges in probability uniformly in R. Our proofs are a slight modification of West (1996), Clark and McCracken (2001) and West and McCracken (1998) and extend their results to weak convergence in the space of functions on $[\underline{\mu},\overline{\mu}]$. The uniform convergence of variance estimators follows from the uniform convergence of second moments of summands in the numerator and the uniform convergence of rolling and recursive estimators, as in the literature on structural change (see Andrews, 1993, for example).

2.1 Non-Nested Model Comparisons

Traditionally, researchers interested in doing inference about the relative forecasting performance of competing, non-nested models rely on the Diebold and Mariano's (1995), West's (1996) and McCracken's (2000) test statistics. The statistic tests the null hypothesis that the expected value of the loss differences evaluated at the pseudo-true parameter values

equals zero. That is, let $\Delta L_T^*(R)$ denote the value of the test statistic evaluated at the true parameter values; then the null hypothesis can be rewritten as: $E\left[\Delta L_T^*(R)\right] = 0$. The test statistic that they propose relies on the sample average of the sequence of standardized out-of-sample loss differences, eq. (1):

$$\Delta L_T(R) \equiv \frac{1}{\widehat{\sigma}_R} P^{-1/2} \sum_{t=R}^{T} \Delta L_{t+h}(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}), \tag{5}$$

where $\widehat{\sigma}_R^2$ is a consistent estimate of the long run variance matrix of the out-of-sample loss differences. A consistent estimate of σ^2 for non-nested model comparisons that does not take into account parameter estimation uncertainty is provided in Diebold and Mariano (1995). Consistent estimates of σ^2 that take into account parameter estimation uncertainty in recursive windows are provided by West (1996) and in rolling and fixed windows are provided by McCracken (2000, p. 203, eqs. 5 and 6). For example, a consistent estimator when parameter estimation error is negligible is:

$$\hat{\sigma}_R^2 = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{t=R}^{T} \Delta L_{t+h}^d \left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) \Delta L_{t+h-i}^d \left(\widehat{\theta}_{t-i,R}, \widehat{\gamma}_{t-i,R}\right), \tag{6}$$

where $\Delta L_{t+h}^d \left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) \equiv \Delta L_{t+h} \left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) - P^{-1} \sum_{t=R}^{T} \Delta L_{t+h} \left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)$ and $q(P)$ is a bandwidth that grows with P (e.g., Newey and West, 1987). In particular, a leading case where (6) can be used is when the same loss function is used for estimation and evaluation. For convenience, we provide the consistent variance estimate for rolling, recursive and fixed estimation schemes in Appendix A in Inoue and Rossi (2012).

Appendix A in Inoue and Rossi (2012) shows that Proposition (1) applies to the test statistic (5) under broad conditions. Examples of typical non-nested models satisfying Proposition 1 (provided that the appropriate moment conditions are satisfied) include linear and non-linear models estimated by any extremum estimator (e.g. Ordinary Least Squares, General Method of Moments and Maximum Likelihood); the data can have serial correlation and heteroskedasticity, but are required to be stationary under the null hypothesis (which rules out unit roots and structural breaks). McCracken (2000) shows that this framework allows for a wide class of loss functions.

Our proposed procedure specialized to two-sided tests of non-nested forecast model comparisons is as follows. Let

$$\mathcal{R}_T = \sup_{R \in \{\underline{R}, \ldots \overline{R}\}} |\Delta L_T(R)|, \tag{7}$$

and

$$\mathcal{A}_T = \frac{1}{\overline{R} - \underline{R} + 1} \sum_{R=\underline{R}}^{\overline{R}} |\Delta L_T(R)|, \tag{8}$$

where $\Delta L_T(R)$ is defined in eq. (5), $R = [\mu T]$, $\underline{R} = [\underline{\mu} T]$, $\overline{R} = [\overline{\mu} T]$, and $\hat{\sigma}^2_R$ is a consistent estimator of σ^2. Reject the null hypothesis $H_0 : \lim_{T\to\infty} E[\Delta L^*_T(R)] = 0$ for all R in favor of the alternative $H_A : \lim_{T\to\infty} E[\Delta L^*_T(R)] \neq 0$ for some R at the significance level α when $\mathcal{R}_T > k^{\mathcal{R}}_\alpha$ or when $\mathcal{A}_T > k^{\mathcal{A}}_\alpha$, where the critical values $k^{\mathcal{R}}_\alpha$ and $k^{\mathcal{A}}_\alpha$ are reported in Table 1, Panel A, for $\underline{\mu} = 0.15$. The critical values of these tests as well as the other tests when $\underline{\mu} = 0.20, 0.25, 0.30, 0.35$ can be found in the working paper version; see Inoue and Rossi (2011).

Researchers might be interested in performing one-sided tests as well. In that case, the tests in eqs. (7) and (8) should be modified follows: $\mathcal{R}_T = \sup_{R\in[\underline{R},\ldots\overline{R}]} \Delta L_T(R)$, $\mathcal{A}_T = \frac{1}{\overline{R}-\underline{R}+1} \sum_{R=\underline{R}}^{\overline{R}} \Delta L_T(R)$. The tests reject the null hypothesis $H_0 : \lim_{T\to\infty} E[\Delta L^*_T(R)] = 0$ for all R in favor of the alternative $H_A : \lim_{T\to\infty} E[\Delta L^*_T(R)] < 0$ for some R at the significance level α when $\mathcal{R}_T > k^{\mathcal{R}}_\alpha$ or when $\mathcal{A}_T > k^{\mathcal{A}}_\alpha$, where the critical values $k^{\mathcal{R}}_\alpha$ and $k^{\mathcal{A}}_\alpha$ are reported in Table 1, Panel B, for $\underline{\mu} = 0.15$.

Finally, it is useful to remind readers that, as discussed in Clark and McCracken (2011b), (5) is not necessarily asymptotically normal even when the models are not nested. For example, when $y_{t+1} = \alpha_0 + \alpha_1 x_t + u_{t+1}$ and $y_{t+1} = \beta_0 + \beta_1 z_t + v_{t+h}$ with x_t independent of z_t and $\alpha_1 = \beta_1 = 0$, the two models are non-nested but (5) is not asymptotically normal. The asymptotic normality result does not hinge on whether or not two models are nested but rather on whether or not the disturbance terms of the two models are numerically identical in population under the null hypothesis.

2.2 Nested Models Comparison

For the case of nested models comparison, we follow Clark and McCracken (2001). Let Model 1 be the parsimonious model, and Model 2 be the larger model that nests Model 1. Let y_{t+h} denote the variable to be forecast and let the period-t forecasts of y_{t+h} from the two models be denoted by $\widehat{y}_{1,t+h}$ and $\widehat{y}_{2,t+h}$: the first ("small") model uses k_1 regressors $x_{1,t}$ and the second ("large") model uses $k_1 + k_2 = k$ regressors $x_{1,t}$ and $x_{2,t}$. Clark and McCracken's

(2001) ENCNEW test is defined as:

$$\Delta L_T^{\mathcal{E}}(R) \equiv P \frac{P^{-1}\sum_{t=R}^{T}\left[\left(y_{t+h}-\widehat{y}_{1,t+h}\right)^2-\left(y_{t+h}-\widehat{y}_{1,t+h}\right)\left(y_{t+h}-\widehat{y}_{2,t+h}\right)\right]}{P^{-1}\sum_{t=R}^{T}\left(y_{t+h}-\widehat{y}_{2,t+h}\right)^2}, \tag{9}$$

where P is the number of out-of-sample predictions available, and $\widehat{y}_{1,t+h}, \widehat{y}_{2,t+h}$ depend on the parameter estimates $\hat{\theta}_{t,R}, \hat{\gamma}_{t,R}$. Note that, since the models are nested, Clark and McCracken's (2001) test is one sided.

Appendix A in Inoue and Rossi (2012) shows that Proposition 1 applies to the test statistic (9) under the same assumptions as in Clark and McCracken (2001). In particular, their assumptions hold for one-step-ahead forecast errors ($h = 1$) from linear, homoskedastic models, OLS estimation, and MSE loss function (as discussed in Clark and McCracken (2001), the loss function used for estimation has to be the same as the loss function used for evaluation).

Our robust procedure specializes to tests of nested forecast model comparisons as follows. Let

$$\mathcal{R}_T^{\mathcal{E}} = \sup_{R\in\{\underline{R},\ldots\overline{R}\}} \Delta L_T^{\mathcal{E}}(R), \tag{10}$$

and

$$\mathcal{A}_T^{\mathcal{E}} = \frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}} \Delta L_T^{\mathcal{E}}(R). \tag{11}$$

Reject the null hypothesis $H_0 : \lim_{T\to\infty} E[\Delta L_T^{\mathcal{E}}(R)] = 0$ for all R at the significance level α against the alternative $H_A : \lim_{T\to\infty} E[\Delta L_T^{\mathcal{E}}(R)] > 0$ for some R when $\mathcal{R}_T^{\mathcal{E}} > k_\alpha^{\mathcal{R}}$ or $\mathcal{A}_T^{\mathcal{E}} > k_\alpha^{\mathcal{A}}$, where the critical values $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$ for $\underline{\mu} = 0.15$ are reported in Table 2.

2.3 Regression-Based Tests of Predictive Ability

Under the widely used MSFE loss, optimal forecasts have a variety of properties. They should be unbiased, one step-ahead forecast errors should be serially uncorrelated, and h-steps-ahead forecast errors should be correlated at most of order $h-1$ (see Granger and Newbold, 1986, and Diebold and Lopez, 1996). It is therefore interesting to test such properties. We do so in the same framework as West and McCracken (1998). Let the forecast error evaluated at the pseudo-true parameter values θ^* be $v_{t+h}(\theta^*) \equiv v_{t+h}$, and its estimated value be $v_{t+h}\left(\widehat{\theta}_{t,R}\right) \equiv \widehat{v}_{t+h}$. We assume one is interested in the linear relationship between the prediction error, v_{t+h}, and a $(p \times 1)$ vector function of data at time t.

For the purposes of this section, let us define the loss function of interest to be $\mathcal{L}_{t+h}(\theta)$, whose estimated counterpart is $\mathcal{L}_{t+h}(\widehat{\theta}_{t,R}) \equiv \widehat{\mathcal{L}}_{t+h}$. To be more specific:

Definition (Special Cases of Regression-based tests of Predictive Ability) *The following are special cases of regression-based tests of predictive ability:*
(i) Forecast Unbiasedness Tests: $\widehat{\mathcal{L}}_{t+h} = \widehat{v}_{t+h}$.
(ii) Mincer-Zarnowitz's (1969) Tests (or Efficiency Tests): $\widehat{\mathcal{L}}_{t+h} = \widehat{v}_{t+h}X_t$, *where* X_t *is a vector of predictors known at time* t *(see also Chao, Corradi and Swanson, 2001). One important special case is when* X_t *is the forecast itself.*
(iii) Forecast Encompassing Tests (Chong and Hendry, 1986, Clements and Hendry, 1993, Harvey, Leybourne and Newbold, 1998): $\widehat{\mathcal{L}}_{t+h} = \widehat{v}_{t+h}f_t$, *where* f_t *is the forecast of the encompassed model.*
(iv) Serial Uncorrelation Tests: $\widehat{\mathcal{L}}_{t+h} = \widehat{v}_{t+h}\widehat{v}_t$.

More generally, let the loss function of interest be the $(p \times 1)$ vector $\mathcal{L}_{t+h}(\theta^*) = v_{t+h}g_t$, whose estimated counterpart is $\widehat{\mathcal{L}}_{t+h} = \widehat{v}_{t+h}\widehat{g}_t$, where $g_t(\theta^*) \equiv g_t$ denotes the function describing the linear relationship between v_{t+h} and a $(p \times 1)$ vector function of data at time t, with $g_t(\widehat{\theta}_t) \equiv \widehat{g}_t$. In the examples above: (i) $g_t = 1$; (ii) $g_t = X_t$; (iii) $g_t = f_t$; (iv) $g_t = v_t$. The null hypothesis of interest is typically:

$$E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0. \tag{12}$$

In order to test (12), one simply tests whether $\widehat{\mathcal{L}}_{t+h}$ has zero mean by a standard Wald test in a regression of $\widehat{\mathcal{L}}_{t+h}$ onto a constant (i.e., testing whether the constant is zero). That is,

$$\mathcal{W}_T(R) = P^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}'_{t+h}\widehat{\Omega}_R^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}, \tag{13}$$

where $\widehat{\Omega}_R$ is a consistent estimate of the long run variance matrix of the adjusted out-of-sample losses, Ω, typically obtained by using West and McCracken's (1998) estimation procedure.

Appendix A in Inoue and Rossi (2012) shows that Proposition 1 applies to the test statistic (13) under broad conditions, which are similar to those discussed for eq. (5). The framework allows for linear and non-linear models estimated by any extremum estimator (e.g. OLS, GMM and MLE), the data to have serial correlation and heteroskedasticity as long as stationary is satisfied (which rules out unit roots and structural breaks), and forecast errors (which can be either one period or multi-period) evaluated using continuously differentiable loss functions, such as MSE.

Our proposed procedure specialized to tests of forecast optimality is the following. Let

$$\mathcal{R}_T^{\mathcal{W}} = \sup_{R\in\{\underline{R},...\overline{R}\}} [\widehat{\mathcal{L}}_T(R)' \widehat{\Omega}_R^{-1} \widehat{\mathcal{L}}_T(R)], \tag{14}$$

and

$$\mathcal{A}_T^{\mathcal{W}} = \frac{1}{\overline{R}-\underline{R}+1} \sum_{R=\underline{R}}^{\overline{R}} [\widehat{\mathcal{L}}_T(R)' \widehat{\Omega}_R^{-1} \widehat{\mathcal{L}}_T(R)], \tag{15}$$

where $\widehat{\mathcal{L}}_T(R) \equiv P^{-1/2}\sum_{t=R}^{T} \widehat{\mathcal{L}}_{t+h}$,and $\widehat{\Omega}_R$ is a consistent estimator of Ω. Reject the null hypothesis $H_0 : \lim_{T\to\infty} E(\mathcal{L}_{t+h}(\theta^*)) = 0$ for all R at the significance level α when $\mathcal{R}_T^{\mathcal{W}} > k_\alpha^{\mathcal{R}}$ for the sup-type test and when $\mathcal{A}_T^{\mathcal{W}} > k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$ for the average-type test, where the critical values $k_{\alpha,p}^{\mathcal{R},\mathcal{W}}$ and $k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$ for $\underline{\mu} = 0.15$ are reported in Table 3.

A simple, consistent estimator for Ω can be obtained by following West and McCracken (1998). West and McCracken (1998) have shown that it is very important to allow for a general variance estimator that takes into account estimation uncertainty and/or correcting the statistics by the necessary adjustments. See West and McCracken's (1998) Table 2 for details on the necessary adjustment procedures for correcting for parameter estimation uncertainty. The same procedures should be implemented to obtain correct inference in regression-based tests in our setup. For convenience, we discuss in detail how to construct a consistent variance estimate in the leading case of Mincer and Zarnowitz's (1969) regressions in Appendix B in Inoue and Rossi (2012) in either rolling, recursive or fixed estimation schemes.

Historically, researchers have estimated the alternative regression: $\widehat{v}_{t+h} = \widehat{g}_t' \cdot \widehat{\alpha}(R) + \widehat{\eta}_{t+h}$, where $\widehat{\alpha}(R) = \left(P^{-1}\sum_{t=R}^{T} \widehat{g}_t \widehat{g}_t'\right)^{-1} \left(P^{-1}\sum_{t=R}^{T} \widehat{g}_t \widehat{v}_{t+h}\right)$ and $\widehat{\eta}_{t+h}$ is the fitted error of the regression, and tested whether the coefficients equal zero. It is clear that under the additional assumption that $E(g_t g_t')$ is full rank (a maintained assumption in that literature) the two procedures share the same null hypothesis and are therefore equivalent. However, in this case it is convenient to define the following re-scaled Wald test:

$$\mathcal{W}_T^{(r)}(R) = \widehat{\alpha}(R)' \ \widehat{V}_\alpha^{-1}(R) \widehat{\alpha}(R),$$

where $\widehat{V}_\alpha(R)$ is a consistent estimate of the asymptotic variance of $\widehat{\alpha}(R)$, V_α. We propose the following tests:

$$\mathcal{R}_T^{\mathcal{W}} = \sup_{R\in\{\underline{R},...\overline{R}\}} \widehat{\alpha}(R)' \ \widehat{V}_\alpha^{-1}(R) \widehat{\alpha}(R), \tag{16}$$

and

$$\mathcal{A}_T^{\mathcal{W}} = \frac{1}{\overline{R}-\underline{R}+1} \sum_{R=\underline{R}}^{\overline{R}} \widehat{\alpha}(R)' \ \widehat{V}_\alpha^{-1}(R) \widehat{\alpha}(R). \tag{17}$$

Reject the null hypothesis $H_0 : \lim_{T\to\infty} E\left[\widehat{\alpha}\left(R\right)\right] = 0$ for all R when $\mathcal{R}_T^{\alpha} > k_{\alpha,p}^{\mathcal{R},\mathcal{W}}$ for the sup-type test and when $\mathcal{A}_T^{\alpha} > k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$ for the average-type test. Simulated values of $k_{\alpha,p}^{\mathcal{R},\mathcal{W}}$ and $k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$ for $\underline{\mu} = 0.15$ and various values of p are reported in Table 3.

Under more general specifications for the loss function, the properties of forecast errors previously discussed may not hold. In those situations, Patton and Timmermann (2007) show that a "generalized forecast error" does satisfy the same properties. The procedures that we propose can also be applied to Patton and Timmermann's (2007) generalized forecast error.

3 Robust Tests of Predictive Accuracy When the Window Size is Small

All the tests considered so far rely on the assumption that the window is a fixed fraction of the total sample size, asymptotically. This assumption rules out the tests by Clark and West (2006, 2007) and Giacomini and White (2005), which rely on a constant (fixed) window size. Propositions 2 and 3 extend our methodology in these two cases by allowing the window size to be fixed.

First, we will consider a version of Clark and West's (2006, 2007) test statistics. Monte Carlo evidence in Clark and West (2006, 2007) and Clark and McCracken (2001, 2005) shows that Clark and West's (2007) test has power broadly comparable to the power of an F-type test of equal MSE. Clark and West's (2006, 2007) test is also popular because it has the advantage of being approximately normal, which permits the tabulation of asymptotic critical values applicable under multi-step forecasting and conditional heteroskedasticity. Before we get into details, a word of caution: our setup requires strict exogeneity of the regressors, which is a very strong assumption in time series application. When the window size diverges to infinity, the correlation between the rolling regression estimator and the regressor vanishes even when the regressor is not strictly exogenous. When the window size is fixed relative to the sample size, however, the correlation does not vanish even asymptotically when the regressor is not strictly exogenous. When the null model is the no-change forecast model as required by the original test of Clark and West (2006, 2007) when the window size is fixed, the assumption of strict exogeneity can be dropped and our test statistic becomes identical to theirs.

Consider the following nested forecasting models:

$$y_{t+h} = \beta_1' x_{1t} + e_{1,t+h}, \tag{18}$$

$$y_{t+h} = \beta_2' x_{2t} + e_{2,t+h}, \tag{19}$$

where $x_{2,t} = [x_{1,t}' \; z_t']'$. Let $\hat{\beta}_{1t}(R) = (\sum_{s=t-R+1}^{t} x_{1,s} x_{1,s}')^{-1} \sum_{s=t-R+1}^{t} x_{1,s} y_{s+h}$ and $\hat{\beta}_{2t}(R) = (\sum_{s=t-R+1}^{t} x_{2,s} x_{2,s}')^{-1} \sum_{s=t-R+1}^{t} x_{2,s} y_{s+h}$ and let $\hat{e}_{1,t+h}(R)$ and $\hat{e}_{2,t+h}(R)$ denote the corresponding models' h-steps-ahead forecast errors. Note that, since the models are nested, Clark and West's (2007) test is one sided. Under the null hypothesis that $\beta_2^* = [\beta_1^{*\prime} \; 0']'$, the *MSPE-adjusted* of Clark and West (2007) can be written as:

$$\begin{aligned} \textit{MSPE-adjusted} &= P^{-1} \sum_{t=R}^{T} \hat{e}_{1,t+h}^2(R) - \left[\hat{e}_{2,t+h}^2(R) - (\hat{y}_{1,t+h} - \hat{y}_{2,t+h})^2\right] \\ &= 2P^{-1} \sum_{t=R}^{T} \hat{e}_{1,t+h}(R) \left[\hat{e}_{1,t+h}(R) - \hat{e}_{2,t+h}(R)\right] \end{aligned}$$

where $P^{-1} \sum_{t=R}^{T} (\hat{y}_{1,t+h} - \hat{y}_{2,t+h})^2$ is the adjustment term. When R is fixed, as Clark and West (2007, p.299) point out, the mean of *MSPE-adjusted* is nonzero unless x_{1t} is null. We consider an alternative adjustment term so that the adjusted loss difference will have zero mean. Suppose that $\beta_2^* = [\beta_1^{*\prime} \; 0']'$ and that $x_{2,t}$ is strictly exogenous. Then we have

$$\begin{aligned} E[(\hat{e}_{1,t+h}^2 - \hat{e}_{2,t+h})^2] &= E[(y_{t+h} - \hat{y}_{1,t+h})^2] - E[(y_{t+h} - \hat{y}_{2,t+h})^2] \\ &= E(y_{t+h}^2 - 2y_{t+h} x_{1t}' \hat{\beta}_{1t} + \hat{y}_{1,t+h}^2) - E(y_{t+h}^2 - 2y_{t+h} x_{2t}' \hat{\beta}_{2t} + \hat{y}_{2,t+h}^2) \\ &= E[\hat{y}_{1,t+h}^2 - \hat{y}_{2,t+h}^2] + 2E[y_{t+h}(x_{2t}' \hat{\beta}_{2t} - x_{1t}' \hat{\beta}_{1t})] \\ &= E[\hat{y}_{1,t+h}^2 - \hat{y}_{2,t+h}^2] + 2E\{y_{t+h}[x_{2t}'(\hat{\beta}_{2t} - \beta_2^*) - x_{1t}'(\hat{\beta}_{1t} - \beta_1^*)]\} \quad (20) \\ &= E[\hat{y}_{1,t+h}^2 - \hat{y}_{2,t+h}^2] + 2E[\beta_1^{*\prime} x_{1t}(x_{2t}'(\hat{\beta}_{2t} - \beta_2^*) - x_{1t}'(\hat{\beta}_{1t} - \beta_1^*))] \quad (21) \\ &= E[\hat{y}_{1,t+h}^2 - \hat{y}_{2,t+h}^2], \end{aligned}$$

where the fourth equality follows from the null hypothesis, $\beta_2^* = [\beta_1^{*\prime} \; 0']'$, the fifth equality follows from the null that $e_{2,t+h}$ is orthogonal to the information set at time t and the last equality from the strict exogeneity assumption. Thus $\phi_{t+h}(R) \equiv \hat{e}_{1,t+h}^2(R) - \hat{e}_{2,t+h}^2(R) - [\hat{y}_{1,t+h}^2(R) - \hat{y}_{2,t+h}^2(R)]$ has zero mean even when x_{1t} is not null provided that the regressors are strictly exogenous.

When R is fixed, Clark and West's adjustment term is valid if the null model is the no-change forecast model, i.e., x_{1t} is null. When x_{1t} is null, the second term on the right-hand side of equation (20) is zero even when x_{2t} is not strictly exogenous, and our adjustment term and theirs become identical.

Proposition 2 (Out-of-Sample Robust Test with Fixed Window Size I.) *Suppose that: (a) either x_{1t} is null or $E(e_{2t}|x_{2s}) = 0$ for all s and t such that $t - \overline{R} \leq s \leq t + \overline{R}$; (b) $\{[e_{1,t+1}, x'_{1,t+1}, z'_{t+1}]'\}$ is α-mixing of size $-r/(r-2)$; (c) $[e_{1,t+h}, x'_{1,t}, z'_t, \hat{\beta}_{1,t}(\underline{R})', \hat{\beta}_{1,t}(\underline{R}+1)', ..., \hat{\beta}_{1,t}(\overline{R}), \hat{\beta}_{2,t}(\underline{R})', \hat{\beta}_{2,t}(\underline{R}+1)', ..., \hat{\beta}_{2,t}(\overline{R})]'$ has finite $4r$-th moments uniformly in t; (d) $\underline{R}$ and $\overline{R}$ are fixed constants. Then*

$$\xi_R \equiv \begin{bmatrix} P^{-1/2} \sum_{t=\overline{R}}^{T} \phi_{t+h}(\underline{R}) \\ P^{-1/2} \sum_{t=\overline{R}+1}^{T} \phi_{t+h}(\underline{R}+1) \\ \vdots \\ P^{-1/2} \sum_{t=\overline{R}}^{T} \phi_{t+h}(\overline{R}) \end{bmatrix} \xrightarrow{d} N(0, \Omega)$$

where Ω is the long-run covariance matrix, $\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j$ and

$$\Gamma_j = E \left\{ \begin{bmatrix} \phi_{t+h}(\underline{R}) \\ \phi_{t+h}(\underline{R}+1) \\ \vdots \\ \phi_{t+h}(\overline{R}) \end{bmatrix} \begin{bmatrix} \phi_{t+h-j}(\underline{R}) \\ \phi_{t+h-j}(\underline{R}+1) \\ \vdots \\ \phi_{t+h-j}(\overline{R}) \end{bmatrix}' \right\}.$$

Let $r = \overline{R} - \underline{R} + 1$. The test that we propose is:

$$CW_T \equiv \xi'_R \widehat{\Omega}^{-1} \xi_R \xrightarrow{d} \chi^2_r, \tag{22}$$

where $\widehat{\Omega}$ is a consistent estimate of Ω. The null hypothesis is rejected at the significance level α for any R when $CW_T > \chi^2_{r;\alpha}$, where $\chi^2_{r;\alpha}$ is the $(1-\alpha)$-th quantile of a chi-square distribution with r degrees of freedom.

The proof of this proposition follows directly from Corollary 24.7 of Davidson (1994, p.387). Assumption (a) is necessary for $\phi_{t+h}(R)$ to have zero mean and is satisfied under the assumption discussed by Clark and West (x_{1t} is not null) or under the assumption that x_{2t} is strictly exogenous. The latter assumption is very strong in the applications of interest.

We also consider the Giacomini and White's (2005) framework. Proposition 3 provides a methodology that can be used to robustify their test for unconditional predictive ability with respect to the choice of the window size.

Proposition 3 (Out-of-sample Robust Tests with Fixed Window Size II.) *Suppose the assumptions of Theorem 4 in Giacomini and White (2005) hold, and that there exists a*

unique window size $R \in \{\underline{R}, ..., \overline{R}\}$ *for which the null hypothesis* $H_0 : \lim_{T\to\infty} E\left[\Delta L_T\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right] = 0$ *holds. Let*

$$GW_T = \inf_{R\in\{\underline{R},...,\overline{R}\}} |\Delta L_T(R)|, \tag{23}$$

where $\Delta L_T(R) \equiv \frac{1}{\widehat{\sigma}_R} T^{-1/2} \sum_{t=R}^{T} \Delta L_T(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R})$, $\underline{R}$ *and* $\overline{R}$ *are fixed constants, and* $\hat{\sigma}_R^2$ *is a consistent estimator of* σ^2*. Under the null hypothesis,*

$$GW_T \xrightarrow{d} N(0,1),$$

The null hypothesis for the GW_T *test is rejected at the significance level* α *in favor of the two-sided alternative* $\lim_{T\to\infty} E\left[\Delta L_T\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right] \neq 0$ *for any* R *when* $GW_T > z_{\alpha/2}$*, where* $z_{\alpha/2}$ *is the* $100(1-\alpha/2)\%$ *quantile of a standard normal distribution.*

Note that, unlike the previous cases, in this case we consider the $inf(\cdot)$ over the sequence of out-of-sample tests rather than the $sup(\cdot)$. The reason why we do so is related to the special nature of Giacomini and White's (2005) null hypothesis: if their null hypothesis is true for one window size then it is necessarily false for other window sizes; thus, the test statistic is asymptotically normal for the former, but diverges for the others. That is why it makes sense to take the $inf(\cdot)$. Our assumption that the null hypothesis holds only for one value of R may sound peculiar, but the unconditional predictive ability test of Giacomini and White (2005) typically implies a unique value of R, although there is no guarantee that the null hypothesis of Giacomini and White (2006) holds in general. For example, consider the the case where data are generated from $y_t = \beta_2^* + e_t$ where $e_t \overset{iid}{\sim} (0, \sigma^2)$, and let the researcher be interested in comparing the MSFE of a model where y_t is unpredictable ($y_t = e_{1t}$) with that of a model where y_t is constant ($y_t = \beta_2 + e_{2,t}$). Under the unconditional version of the null hypothesis we have $E[y_{t+1}^2 - (y_{t+1} - R^{-1}\Sigma_{j=t-R+1}^{t} y_j)^2] = 0$, which in turn implies $\beta_2^{*2} - \frac{\sigma^2}{R} = 0$. Thus, if the null hypothesis holds then it holds with a unique value of R. Our proposed test protects applied researchers from incorrectly rejecting the null hypothesis by choosing an ad hoc window size, which is important especially for the Giacomini and White's (2005) test, given its sensitivity to data snooping over window sizes.

The proof of Proposition 3 is provided in Appendix A in Inoue and Rossi (2012). Note that one might also be interested in a one-sided test, where $H_0 : \lim_{T\to\infty} E\left[\Delta L_T\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right] = 0$ versus the alternative that $\lim_{T\to\infty} E\left[\Delta L_T\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right] > 0$. In that case, construct $GW_T = \inf_{R=\underline{R},...\overline{R}} \Delta L_T(R)$, and reject when $GW_T > z_\alpha$, where $z_{\alpha/2}$ is the $100(1-\alpha)\%$ quantile of a standard normal distribution.

4 Monte Carlo evidence

In this section, we evaluate the small sample properties of the methods that we propose and compare them with the methods existing in the literature. We consider both nested and non-nested models' forecast comparisons, as well as forecast rationality. For each of these tests under the null hypothesis, we allow for three choices of $\underline{\mu}$, one-step-ahead and multi-step-ahead forecasts, and multiple regressors of alternative models to see if and how the size of the proposed tests is affected in small samples. We consider the no-break alternative hypothesis and the one-time-break alternative to compare the power of our proposed tests with that of the conventional tests. Below we report rejection frequencies at the 5% nominal significance level to save space.

For the nested models comparison, we consider a modification of the DGP (labeled "DGP 1") that follows Clark and McCracken (2005a) and Pesaran and Timmermann (2007). Let

$$\begin{pmatrix} y_{t+1} \\ x_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} 0.3 & d_{t,T} & 0_{1\times(k_2-1)} \\ 0 & 0.5 & 0_{1\times(k_2-1)} \\ 0 & 0_{(k_2-1)\times 1} & 0.5 \cdot I_{k_2-1} \end{pmatrix} \begin{pmatrix} y_t \\ x_t \\ z_t \end{pmatrix} + \begin{pmatrix} u_{y,t+1} \\ u_{x,t+1} \\ u_{z,t+1} \end{pmatrix}, \quad t = 1, ..., T-1,$$

where $y_0 = x_0 = 0$, $z_0 = 0_{(k_2-1)\times 1}$, $[u_{y,t+1} \;\; u_{x,t+1} \;\; u'_{z,t+1}]' \overset{iid}{\sim} N(0_{(k_2+1)\times 1}, I_{k_2+1})$ and I_{k_2+1} denotes an identity matrix of dimension $(k_2+1)\times(k_2+1)$. We compare the following two nested models' forecasts for y_{t+h}:

$$\begin{aligned} &\text{Model 1 forecast} \;\; : \;\; \widehat{\theta}_{1,t} y_t \\ &\text{Model 2 forecast} \;\; : \;\; \widehat{\gamma}_{1,t} y_t + \widehat{\gamma}'_{2,t} x_t + \widehat{\gamma}'_{3,t} z_t, \end{aligned} \tag{24}$$

and both models' parameters are estimated by OLS in rolling windows of size R. Under the null hypothesis $d_{t,T} = 0$ for all t and we consider $h = 1, 4, 8$, $k_2 = 1, 3, 5$ and $T = 50, 100, 200, 500$. We consider several horizons (h) to evaluate how our tests perform at both the short and long horizons that are typically considered in the literature. We consider several extra-regressors (k_2) to evaluate how our tests perform as the estimation uncertainty induced by extra regressors increases. Finally, we consider several sample sizes (T) to evaluate how our tests perform in small samples. Under the no-break alternative hypothesis $d_{t,T} = 0.1$ or $d_{t,T} = 0.2$ ($h = 1$, $k_2 = 1$ and $T = 200$). Under the one-time-break alternative hypothesis, $d_{t,T} = 0.5 \cdot I(t \leq \tau)$ for $\tau \in \{40, 80, 120, 160\}$, ($h = 1$, $k_2 = 1$ and $T = 200$).

For the non-nested models' comparison, we consider a modification of DGP1 (labeled

"DGP2"):

$$\begin{pmatrix} y_{t+1} \\ x_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} 0.3 & d_{t,T} & 0.5 \;\; 0_{1\times(k-2)} \\ 0 & 0.5 & 0 \;\; 0_{1\times(k-2)} \\ 0_{(k-1)\times 1} & 0_{(k-1)\times 1} & 0.5I_{(k-1)} \end{pmatrix} \begin{pmatrix} y_t \\ x_t \\ z_t \end{pmatrix} + \begin{pmatrix} u_{y,t+1} \\ u_{x,t+1} \\ u_{z,t+1} \end{pmatrix}, \;\; t = 1, ..., T-1,$$

where $y_0 = x_0 = 0$, $z_0 = 0_{(k-1)\times 1}$, and $[u_{y,t+1} \; u_{x,t+1} \; u'_{z,t+1}]' \overset{iid}{\sim} N(0_{(k+1)\times 1}, I_{k+1})$. We compare the following two non-nested models' forecasts for y_{t+h}:

$$\begin{aligned} \text{Model 1 forecast} \;\; &: \;\; \widehat{\theta}_1 y_t + \widehat{\theta}_2 x_t \\ \text{Model 2 forecast} \;\; &: \;\; \widehat{\gamma}_1 y_t + \widehat{\gamma}'_2 z_t, \end{aligned} \tag{25}$$

and both models' parameters are estimated by OLS in rolling windows of size R. Under the null hypothesis $d_{t,T} = 0.5$ for all t. Again, we consider several horizons, number of extra-regressors and sample sizes: $h = 1, 4, 8$, $k = 2, 4, 6$ and $T = 50, 100, 200, 500$. We use the two-sided version of our test. Note that, for non-nested models with $k > 2$, one might expect that, in finite samples, model 1 would be more accurate than model 2 because model 2 includes extraneous variables, however. Under the no-break alternative hypothesis $d_{t,T} = 1$ or $d_{t,T} = 1.5$ ($h = 1$, $k = 2$ and $T = 200$). Under the one-time-break alternative hypothesis, $d_{t,T} = 0.5 \cdot I(t \leq \tau) + 0.5$ for $\tau \in \{40, 80, 120, 160\}$, ($h = 1$, $k = 2$ and $T = 200$).

"DGP3" is designed for regression-based tests and is a modification of the Monte Carlo design in West and McCracken (1998). Let

$$y_{t+1} \;=\; \delta_{t,T} \cdot I_p + 0.5 y_t + \varepsilon_{t+1}, \;\; t = 1, ..., T,$$

where y_{t+1} is a $p \times 1$ vector and $\varepsilon_{t+1} \overset{iid}{\sim} N(0_{p\times 1}, I_p)$. We generate a vector of variables rather than a scalar because in this design we are interested in testing whether the forecast error is not only unbiased but also uncorrelated with information available up to time t, including lags of the additional variables in the model. Let $y_{1,t}$ be the first variable in the vector y_t. We estimate $y_{1,t+h} \;=\; \theta' y_t + v_{t+h}$ by rolling regressions and test $E(v_{t+h}) = 0$ and $E(y_t v_{t+h}) = 0$ for $h = 1, 4, 8$ and $p = 1, 3, 6$. We let $\delta_{t,T} = 0.5$ or $\delta_{t,T} = 1$ under the no-break alternative and $\delta_{t,T} = 0.5 \cdot I(t \leq \tau)$ for $\tau \in \{40, 80, 120, 160\}$ under the one-time break alternative ($h = 1$, $p = 1$ and $T = 200$).

For the forecast comparison tests with a fixed window size, we consider the following DGP (labeled "DGP4"): $y_{t+1} \;=\; \delta_R x_t + \varepsilon_{t+1}, \;\; t = 1, ..., T$,where x_t and ε_{t+1} are i.i.d. standard Normal independent of each other. We compare the following two nested models'

forecasts for y_t: a first model is a no-change forecast model, e.g. the random walk forecast for a target variable defined in first differences, and the second is a model with the regressor; that is, Model 1 forecast equals zero and Model 2 forecast equals $\widehat{\delta}_{t,R}x_t$,where $\widehat{\delta}_{t,R} = \left(\sum_{j=t-R+1}^{t} x_t^2\right)^{-1} \sum_{j=t-R+1}^{t} x_t y_t$. To ensure that the null hypothesis in Proposition 3 holds for one of the window sizes, $\underline{R}$, we let $\delta_R = (\underline{R}-2)^{-1/2}$. The number of Monte Carlo replications is 5,000. To ensure that the null hypothesis in Proposition 2 holds, we let $\delta_R = \delta = 0$.

The size properties of our test procedures in small samples are first evaluated in a series of Monte Carlo experiments. We report empirical rejection probabilities of the tests we propose at the 5% nominal level. In all experiments except DGP4, we investigate sample sizes where $T = 50$, 100, 200 and 500 and set $\underline{\mu} = 0.05, 0.15, 0.25$ and $\overline{\mu} = 1 - \underline{\mu}$. For DGP4, we let $P = 100$, 200 and 500 and let $\underline{R} = 20$ or 30 and $\overline{R} = \underline{R} + 5$. Note that in design 4 we only consider five values of R since the window size is small by assumption, and that limits the range of values we can consider. Tables 4, 5 and 6 report results for the $\mathcal{R}_T^{\varepsilon}$ and $\mathcal{A}_T^{\varepsilon}$ tests for the nested models comparison (DGP1), the $\mathcal{R}_T$ and $\mathcal{A}_T$ tests for non-nested models comparison (DGP2), and $\mathcal{R}_T^W$ and $\mathcal{A}_T^W$ for the regression-based tests of predictive ability (DGP3), respectively. For the multiple horizon case, in nested and regression-based inference we use the heteroskedasticity and autocorrelation consistent (HAC) estimator with the truncated kernel, bandwidth $h-1$ and the adjustment proposed by Harvey, Leybourne and Newbold (1997), as suggested by Clark and McCracken (2011a, Section 4), and then bootstrap the test statistics using the parametric bootstrap based on the estimated VAR model as suggested by Clark and McCracken (2005). Note that designs that have the same parameterization do not have exactly the same rejection frequencies since the Monte Carlo experiments are ran independently for the various cases we study, and therefore there are small differences due to simulation noise. The number of Monte Carlo simulations is set to 5,000 except that it is set to 500 and the number of bootstrap replications is 199 in Tables 4 and 6 when $h > 1$.

Table 4 shows that the nested model comparison tests (i.e., $\mathcal{R}_T^{\mathcal{E}}$ and $\mathcal{A}_T^{\mathcal{E}}$ tests) have good size properties overall. Except for small sample sizes, they perform well even in the multiple forecast horizon and multiple regressor cases. Although the effect of the choice of $\underline{\mu}$ becomes smaller as the sample size grows, the $\mathcal{R}_T^{\mathcal{E}}$ test tends to over-reject with smaller values of $\underline{\mu}$. The $\mathcal{A}_T^{\mathcal{E}}$ test is less sensitive to the choice of $\underline{\mu}$. The tests implemented with $\underline{\mu} = 0.05$ tend to reject the null hypothesis too often when the sample size is small. For the size properties we

recommend that $\underline{\mu} = 0.15$. Table 5 shows that the non-nested model comparison tests ($\mathcal{R}_T$ and $\mathcal{A}_T$ tests) also have good size properties although they tend to be slightly under-sized. They tend to be more under-sized as the forecast horizon grows, thus suggesting that the test is less reliable for horizons greater than one period. The $\mathcal{R}_T$ test tends to reject too often when there are many regressors ($p = 6$). Note that, by showing that the test is significantly oversized in small samples, the simulation results confirm that for non-nested models with $p > 1$ model 1 should be more accurate than model 2 in finite samples, as expected. Table 6 shows the size properties of the regression-based tests of predictive ability ($\mathcal{R}_T^W$ and $\mathcal{A}_T^W$ tests). The tests tend to reject more often as the the forecast horizon increases and less often as the number of restrictions increases.

Table 7 reports empirical rejection frequencies for DGP4. The left panel shows results for the GW_T test, eq. (23), reported in the column labeled "GW_T test". The table shows that our test is conservative when the number of out-of-sample forecasts P is small, but otherwise it is controlled. Similar results hold for the CW_T test discussed in Proposition 2.

Next, we consider three additional important issues. First, we evaluate the power properties of our proposed procedure in the presence of departures from the null hypothesis in small samples. Second, we show that traditional methods, which rely on an "ad-hoc" window size choice, may have no power at all to detect predictive ability. Third, we demonstrate traditional methods are subject to data mining (i.e. size distortions) if they are applied to many window sizes without correcting the appropriate critical values.

Tables 8, 9 and 10 report empirical rejection rates for the Clark and McCracken's (2001) test under DGP1 with $h = 1$ and $p = 0$, the non-nested model comparison test of Diebold and Mariano (1995), West (1996) and McCracken (2000) under DGP2 with $h = 1$ and $p = 1$, and West and McCracken's (1998) regression-based test of predictive ability under DGP3 with $h = 1$ and $p = 1$, respectively. In each table, the columns labeled "Tests Based on Single R" report empirical rejection rates implemented with a specific value of R which would correspond to the case of a researcher who has chosen one "ad-hoc" window size R, has not experimented with other choices, and thus might have missed predictive ability associated with alternative values of R. The columns labeled "Data Mining" report empirical rejection rates incurred by a researcher who is searching across all values of $R \in \{30, 31, ..., 170\}$ ("all R") and across five values, $R \in \{20, 40, 80, 120, 160\}$. That is, the researcher reports results associated with the most significant window size without taking into account the search procedure when doing inference. The critical values used for these conventional testing procedures are based on Clark and McCracken (2001) and West and McCracken (1998)

for Tables 8 and 10 and are equal to 1.96 for Table 9. Note that to obtain critical values for the ENCNEW test and regression-based test of predictive ability that are not covered by their tables, the critical values are estimated from 50,000 Monte Carlo simulations in which the Brownian motion is approximated by normalized partial sums of 10,000 standard normal random variates. For the non-nested model comparison test, parameter estimation uncertainty is asymptotically irrelevant by construction and the standard normal critical values can be used. The nominal level is set to 5%, $\underline{\mu} = 0.15$, $\overline{\mu} = 0.85$, and the sample size is 200.

The first row of each panel reports the size of these testing procedures and shows that all tests have approximately the correct size except the data mining procedure, which has size distortions and leads to too many rejections with probabilities ranging from 0.175 to 0.253. Even when only five window sizes are considered, data mining leads to falsely rejecting the null hypothesis with probability more than 0.13. This implies that the empirical evidence in favor of the superior predictive ability of a model can be spurious if evaluated with the incorrect critical values. Results in Inoue and Rossi (2012, Panel A in Tables 8, 9 and 10) show that the conventional tests and proposed tests have power against the standard no-break alternative hypothesis. Unreported results show that while the power of the $\mathcal{R}_T^{\mathcal{E}}$ test is increasing in $\underline{\mu}$, it is decreasing in $\underline{\mu}$ for the $\mathcal{R}_T$ and $\mathcal{R}_T^W$ tests. The power of the $\mathcal{A}_T^{\mathcal{E}}$, $\mathcal{A}_T$ and $\mathcal{A}_T^W$ tests is not sensitive to the choice of $\underline{\mu}$.

The tables demonstrate that, in the presence of a structural break the tests based on an "ad-hoc" rolling window size can have low power depending on the window size and the break location. The evidence highlights the sharp sensitivity of power of all the tests to the timing of the break relative to the forecast evaluation window, and shows that, in the presence of instabilities, our proposed tests tend to be more powerful than some of the tests based on an ad-hoc window size, whose power properties crucially depend on the window size. Against the break alternative, the power of the proposed tests tend to be decreasing in $\underline{\mu}$. Based on these size and power results we recommend $\underline{\mu} = 0.15$ in Section 2, which provides a good performance overall.

Finally, we show that the effects of data mining are not just a small sample phenomenon. We quantify the effects of data mining asymptotically by using the limiting distributions of existing test statistics. We design a Monte Carlo simulation where we generate a large sample of data (T=2000) and use it to construct limiting approximations to the test statistics described in Appendix B in Inoue and Rossi (2012). For example, in the non-nested models comparison case with $p = 1$, the limiting distribution of the Diebold and Mariano (1995) test

statistic for a given $\mu = \underset{T\to\infty}{lim} \frac{R}{T}$ is $(1-\mu)^{-1/2} |B(1) - B(\mu)|$; the latter can be approximated in large samples by $\left(1-\frac{R}{T}\right)^{-1/2} |P^{-1/2} \sum_{t=R}^{T} \xi_t|$, where $\xi_t \sim iidN(0,1)$. We simulate the latter for many window sizes R and then calculate how many times, on average across 50,000 Monte Carlo replications, the resulting vector of statistics exceed the standard normal critical values for a 5% nominal size. Table 11 reports the results, which demonstrate that the over-rejections of traditional tests when researchers data snoop over window sizes persist asymptotically.

5 Empirical evidence

The poor forecasting ability of economic models of exchange rate determination has been recognized since the works by Meese and Rogoff (1983a,b), who established that a random walk forecasts exchange rates better than any economic models in the short run. Meese and Rogoff's (1983a,b) finding has been confirmed by several researchers and the random walk is now the yardstick of comparison for the evaluation of exchange rate models. Recently, Engel, Mark and West (2007) and Molodtsova and Papell (2009) documented empirical evidence in favor of the out-of-sample predictability of some economic models, especially those based on the Taylor rule. However, the out-of-sample predictability that they report depends on certain parameters, among which the choice of the in-sample and out-of-sample periods and the size of the rolling window used for estimation. The choice of such parameters may affect the outcome of out-of-sample tests of forecasting ability in the presence of structural breaks. Rossi (2006) found empirical evidence of instabilities in models of exchange rate determination; Giacomini and Rossi (2010) evaluated the consequences of instabilities in the forecasting performance of the models over time; Rogoff and Stavrakeva (2008) also question the robustness of these results to the choice of the starting out-of-sample period. In this section, we test the robustness of these results to the choice of the rolling window size. It is important to notice that it is not clear a-priori whether our test would find more or less empirical evidence in favor of predictive ability. In fact, there are two opposite forces at play. By considering a wide variety of window sizes, our tests might be *more* likely to find empirical evidence in favor of predictive ability, as our Monte Carlo results have shown. However, by correcting statistical inference to take into account the search process across multiple window sizes, our tests might at the same time be *less* likely to find empirical evidence in favor of predictive ability.

Let s_t denote the logarithm of the bilateral nominal exchange rate, where the exchange rate is defined as the domestic price of foreign currency. The rate of growth of the exchange rate depends on its deviation from the current level of a macroeconomic fundamental. Let f_t denote the long-run equilibrium level of the nominal exchange rate as determined by the macroeconomic fundamental, and $z_t = f_t - s_t$. Then,

$$s_{t+1} - s_t = \alpha + \beta z_t + \varepsilon_{t+1} \tag{26}$$

where ε_{t+1} is an unforecastable error term.The first model we consider is the Uncovered Interest Rate Parity (UIRP). In the UIRP model,

$$f_t^{UIRP} = (i_t - i_t^*) + s_t, \tag{27}$$

where $(i_t - i_t^*)$ is the short-term interest differential between the home and the foreign countries.

The second model we consider is a model with Taylor rule fundamentals, as in Molodtsova and Papell (2009) and Engel, Mark and West (2007). Let π_t denote the inflation rate in the home country, π_t^* denote the inflation rate in the foreign country, $\overline{\pi}$ denote the target level of inflation in each country, y_t^{gap} denote the output gap in the home country and y_t^{gap*} denote the output gap in the foreign country. Note that the output gap is the percentage difference between actual and potential output at time t, where the potential output is the linear time trend in output, and that Taylor rule specification is one for which Papell and Molodtsova (2009) find the least empirical evidence of predictability so our results can be interpreted as a lower bound on the predictability of Taylor rules that they consider. Since the difference in the Taylor rule of the home and foreign countries implies $i_t - i_t^* = \delta(\pi_t - \pi_t^*) + \gamma(y_t^{gap} - y_t^{gap*})$, we have that the latter determines the long run equilibrium level of the nominal exchange rate:

$$f_t^{TAYLOR} = \delta(\pi_t - \pi_t^*) + \gamma(y_t^{gap} - y_t^{gap*}) + s_t. \tag{28}$$

The benchmark model, against which the forecasts of both models (27) and (28) are evaluated, is the random walk, according to which the exchange rate changes are forecast to be zero. We chose the random walk without drift to be the benchmark model because it is the toughest benchmark to beat (see Meese and Rogoff, 1983a,b). We use monthly data from the International Financial Statistics database (IMF) and from the Federal Reserve Bank of St. Louis from 1973:3 to 2008:1 for Japan, Switzerland, Canada, Great Britain, Sweden, Germany, France, Italy, the Netherlands, and Portugal. Data on interest rates were incomplete for Portugal and the Netherlands, so we do not report UIRP results for these

countries. The former database provides the seasonally adjusted industrial production index for output, and the 12-month difference of the CPI for the annual inflation rate, and the interest rates. The latter provides the exchange rate series. The two models' rolling forecasts (based on rolling windows calculated over an out-of-sample portion of the data starting in 1983:2) are compared to the forecasts of the random walk, as in Meese and Rogoff (1983a,b). We focus on the methodologies in Section 2.2 since the models are nested. In our exercise, $\underline{\mu} = 0.15$, which implies $\overline{R} = \underline{\mu}T$ and $\underline{R} = (1 - \underline{\mu})T$; the total sample size T depends on the country, and the values of $\overline{R}$ and $\underline{R}$ are shown on the x-axes in Figures 1 and 2, and offer a relatively large range of window sizes, all of which are sufficiently large for asymptotic theory to provide a good approximation.

Empirical results for selected countries are shown in Table 12 and Figure 1 (see Inoue and Rossi, 2012, for detailed results on other countries/models). The column labeled "Test Based on Single R" in Table 12 reports the empirical results in the literature based on a window size R equal to 120, the same window size used in Molodtsova and Papell (2009). According to the "Test Based on Single R," the Taylor model significantly outperforms a random walk for Canada and the U.K. at 5% significance level, whereas the UIRP model outperforms the random walk for Canada and Italy at the 5% significance level. According to our tests, instead, the empirical evidence in favor of predictive ability is much more favorable. Figure 1 reports the estimated Clark and McCracken's (2001) test statistic for the window sizes we consider for the UIRP model. Note that the $\mathcal{R}^{\varepsilon}_{T}$ test rejects if, for any window size R (reported on the x-axis), the test statistic is above the critical value line (dotted lines). In particular, the predictive ability of the economic models tends to show up at smaller window sizes, as the figures show. This suggests that the empirical evidence in favor of predictive ability may be driven by the existence of instabilities in the predictive ability, for which rolling windows of small size are advantageous. One should also be aware of the possibility of data snooping over country-model pairs; we refer to Molodtsova and Papell (2009).

6 Conclusions

This paper proposes new methodologies for evaluating economic models' forecasting performance that are robust to the choice of the estimation window size. These methodologies are noteworthy since they allow researchers to reach empirical conclusions that do not depend on a specific estimation window size. We show that tests traditionally used by forecasters suffer from size distortions if researchers report, in reality, the best empirical result over various

window sizes, but without taking into account the search procedure when doing inference in practice. Traditional tests may also lack power to detect predictive ability when implemented for an "ad-hoc" choice of the window size. Finally, our empirical results demonstrate that the recent empirical evidence in favor of exchange rate predictability is even stronger when allowing a wider search over window sizes.

References

[1] Andrews, D.W.K. (1993), "Tests of Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–856.

[2] Billingsley, P. (1968), *Convergence of Probability Measures,* John Wiley & Sons: New York, NY.

[3] Chao, J.C., V. Corradi and N.R. Swanson (2001), "An Out-of-Sample Test for Granger Causality," *Macroeconomic Dynamics*.

[4] Cheung, Y., M.D. Chinn and A.G. Pascual (2005), "Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive?," *Journal of International Money and Finance* 24, 1150-1175.

[5] Chinn, M. (1991), "Some Linear and Nonlinear Thoughts on Exchange Rates," *Journal of International Money and Finance* 10, 214-230.

[6] Chong, Y.Y. and D.F. Hendry (1986), "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies* 53, 671-690.

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

[8] Clark, T.E. and M.W. McCracken (2005a), "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics* 124, 1-31.

[9] Clark, T.E. and M.W. McCracken (2005b), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24(3), 369-404.

[10] Clark, T.E. and M.W. McCracken (2009), "Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts," *International Economic Review*, 50(2), 363-395.

[11] Clark, T.E. and M.W. McCracken (2010), "Reality Checks and Nested Forecast Model Comparisons," *mimeo*, St. Louis Fed.

[12] Clark, T.E. and M.W. McCracken (2011a), "Advances in Forecast Evaluation," in: G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 2, Elsevier, forthcoming.

[13] Clark, T.E. and M.W. McCracken (2011b), "Tests of Equal Forecast Accuracy for Overlapping Models," *Federal Reserve Bank of St. Louis Working Paper* 2011-024,.

[14] Clark, T.E. and K.D. West (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* 135, 155–186.

[15] Clark, T.E. and K.D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

[16] Clements, M.P. and D.F. Hendry (1993), "On the Limitations of Comparing Mean Square Forecast Errors," *Journal of Forecasting* 12, 617-637.

[17] Davidson, J. (1994), *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

[18] Diebold, F.X. and J. Lopez (1996), "Forecast Evaluation and Combination," in *Handbook of Statistics*, G.S. Maddala and C.R. Rao eds., North-Holland, 241–268.

[19] Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics,* 13, 253-263.

[20] Engel, C., N. Mark and K.D. West, "Exchange Rate Models Are Not as Bad as You Think," in: *NBER Macroeconomics Annual,* Daron Acemoglu, Kenneth S. Rogoff and Michael Woodford, eds. (Cambridge, MA: MIT Press, 2007).

[21] Giacomini, R. and B. Rossi (2010), "Model Comparisons in Unstable Environments", *Journal of Applied Econometrics* 25(4)*,* 595-620.

[22] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability," *Econometrica,* 74, 1545-1578.

[23] Gourinchas, P.O., and H. Rey (2007), "International Financial Adjustment," *The Journal of Political Economy* 115(4), 665-703.

[24] Granger, C.W.J. and P. Newbold (1986), *Forecasting Economic Time Series* (2nd ed.), New York: Academic Press.

[25] Hansen, B.E. (1992), "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes," *Econometric Theory* 8, 489-500.

[26] Hansen, P.R. and A. Timmermann (2011), "Choice of Sample Split in Out-of-Sample Forecast Evaluation", *mimeo.*

[27] Harvey, D.I., S.J. Leybourne and P. Newbold (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting,* 13, 281–291.

[28] Harvey, D.I., S.J. Leybourne and P. Newbold (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16 (2), 254-259.

[29] Inoue, A. and B. Rossi (2011), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size", *ERID Working Paper*, Duke University.

[30] Inoue, A. and B. Rossi (2012), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size", *JBES website.*

[31] McCracken, M.W. (2000), "Robust Out-of-Sample Inference," *Journal of Econometrics* 99, 195-223.

[32] Meese, R. and K.S. Rogoff (1983a), "Exchange Rate Models of the Seventies. Do They Fit Out of Sample?," *The Journal of International Economics* 14, 3-24.

[33] Meese, R. and K.S. Rogoff (1983b), "The Out of Sample Failure of Empirical Exchange Rate Models," in Jacob Frankel (ed.), *Exchange Rates and International Macroeconomics*, Chicago: University of Chicago Press for NBER.

[34] Mincer, J. and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations*, ed. J. Mincer, New York: National Bureau of Economic Research, 81–111.

[35] Molodtsova, T. and D.H. Papell (2009), "Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals," *Journal of International Economics* 77(2).

[36] Newey, W. and K.D. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703-708.

[37] Patton, A.J. and A. Timmermann (2007), "Properties of Optimal Forecasts Under Asymmetric Loss and Nonlinearity," *Journal of Econometrics* 140, 884-918.

[38] Paye, B. and A. Timmermann (2006), "Instability of Return Prediction Models," *Journal of Empirical Finance* 13(3), 274-315.

[39] Pesaran, M.H., D. Pettenuzzo and A. Timmermann (2006), "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies* 73, 1057-1084.

[40] Pesaran, M.H. and A. Timmermann (2005), "Real-Time Econometrics," *Econometric Theory* 21(1), pages 212-231.

[41] Pesaran, M.H. and A. Timmermann (2007), "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics* 137(1), 134-161.

[42] Qi, M. and Y. Wu (2003), "Nonlinear Prediction of Exchange Rates with Monetary Fundamentals," *Journal of Empirical Finance* 10, 623-640.

[43] Rogoff, K.S. and V. Stavrakeva, "The Continuing Puzzle of Short Horizon Exchange Rate Forecasting," *NBER Working paper* No. 14071, 2008.

[44] Rossi, B. (2006), "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability," *Macroeconomic Dynamics* 10(1), 20-38.

[45] Rossi, B., and T. Sekhposyan (2010), "Understanding Models' Forecasting Performance," *mimeo*, Duke University.

[46] Rossi, B., and A. Inoue (2011), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size," mimeo, Duke University and North Carolina State University.

[47] Stock, J.H. and M.W. Watson (2003a), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature.*

[48] West, K.D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067-1084.

[49] West, K.D., and M.W. McCracken (1998), "Regression-Based Tests of Predictive Ability," *International Economic Review,* 39, 817–840.

Tables and Figures

Table 1. Critical Values for Non-Nested Model Comparisons

	Panel A. Two-Sided Critical Values						**Panel B. One-Sided Critical Values**					
	R_T test			A_T test			R_T test			A_T test		
$\underline{\mu}$	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
0.15	2.465	2.754	3.337	1.462	1.739	2.292	2.127	2.458	3.106	1.134	1.454	2.073
0.20	2.398	2.697	3.282	1.489	1.771	2.345	2.057	2.399	3.059	1.158	1.488	2.116
0.25	2.333	2.641	3.228	1.512	1.809	2.394	1.986	2.332	3.007	1.179	1.510	2.167
0.30	2.264	2.577	3.159	1.539	1.838	2.433	1.920	2.261	2.953	1.201	1.535	2.205
0.35	2.186	2.498	3.099	1.564	1.864	2.475	1.838	2.186	2.862	1.225	1.560	2.240

Notes to Table 1. $\underline{\mu}$ is the fraction of the smallest window size relative to T, $\underline{\mu} = \lim_{T\to\infty}(\underline{R}/T)$. The critical values are obtained by Monte Carlo simulation using 50,000 replications, approximating Brownian motions by normalized partial sums of 10,000 standard normals.

Table 2. Critical Values for Nested Model Comparisons Using ENCNEW

	Rolling Regressions						**Recursive Regressions**					
	$\mathcal{R}_T^{\mathcal{E}}$ test			$\mathcal{A}_T^{\mathcal{E}}$ test			$\mathcal{R}_T^{\mathcal{E}}$ test			$\mathcal{A}_T^{\mathcal{E}}$ test		
k_2	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	3.938	5.210	8.124	1.060	1.721	3.434	2.042	3.063	5.620	0.862	1.455	2.861
2	5.623	7.194	10.710	1.602	2.446	4.370	3.122	4.313	7.243	1.315	2.019	3.644
3	6.908	8.676	12.614	2.036	2.987	5.015	3.854	5.200	8.406	1.662	2.427	4.194
4	7.941	9.980	14.451	2.376	3.373	5.597	4.508	5.975	9.501	1.916	2.789	4.701
5	8.892	11.089	15.748	2.665	3.763	6.074	5.050	6.602	10.227	2.165	3.072	5.172
6	9.703	12.029	17.131	2.901	4.074	6.632	5.575	7.201	11.009	2.370	3.331	5.438
7	10.466	12.968	18.405	3.151	4.388	7.029	6.037	7.795	11.737	2.543	3.621	5.755
8	11.225	13.831	19.489	3.360	4.671	7.432	6.476	8.305	12.386	2.731	3.852	6.152
9	11.888	14.585	20.525	3.554	4.946	7.807	6.894	8.829	12.984	2.932	4.102	6.436
10	12.502	15.408	21.415	3.728	5.179	8.172	7.292	9.240	13.598	3.065	4.292	6.700
11	13.105	16.098	22.365	3.903	5.386	8.552	7.614	9.581	14.198	3.210	4.473	7.002
12	13.728	16.787	23.404	4.079	5.614	8.893	7.942	10.075	14.636	3.350	4.646	7.276

Notes. k_2 is the number of additional regressors in the nesting model. Critical values are obtained by 50,000 Monte Carlo simulations, approximating Brownian motions by normalized partial sums of 10,000 std. normals.

Table 3. Critical Values for Regression-Based Forecasts Tests

	Rolling Regressions						**Recursive Regressions**					
	$\mathcal{R}_T^{\mathcal{W}}$ test			$\mathcal{A}_T^{\mathcal{W}}$ test			$\mathcal{R}_T^{\mathcal{W}}$ test			$\mathcal{A}_T^{\mathcal{W}}$ test		
p	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	6.023	7.556	10.909	2.335	3.192	5.291	0.866	0.871	0.880	0.511	0.514	0.520
2	8.710	10.376	14.182	4.032	5.103	7.668	1.723	1.730	1.743	1.015	1.020	1.028
3	10.917	12.773	16.559	5.596	6.796	9.543	2.579	2.587	2.602	1.519	1.524	1.535
4	12.889	14.832	19.144	6.982	8.296	11.240	3.433	3.443	3.460	2.022	2.028	2.040
5	14.717	16.761	21.272	8.394	9.798	12.859	4.287	4.298	4.318	2.525	2.532	2.545
6	16.506	18.595	23.201	9.685	11.249	14.450	5.141	5.152	5.174	3.027	3.035	3.050
7	18.207	20.397	25.005	10.987	12.604	15.961	5.994	6.007	6.031	3.529	3.538	3.554
8	19.860	22.116	26.943	12.238	13.937	17.591	6.847	6.861	6.887	4.031	4.041	4.057
9	21.452	23.852	28.776	13.519	15.291	19.063	7.700	7.714	7.742	4.533	4.543	4.561
10	23.021	25.479	30.488	14.770	16.569	20.384	8.552	8.567	8.597	5.035	5.045	5.065
11	24.474	27.037	32.404	15.955	17.854	21.819	9.405	9.421	9.451	5.537	5.548	5.567
12	26.018	28.607	34.135	17.151	19.131	23.354	10.257	10.274	10.305	6.039	6.049	6.070

Notes. p is the number of restrictions. The critical values are obtained by Monte Carlo simulation using 50,000 replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normals.

Table 4. Size of Nested Model Comparison Tests — DGP1

	$\mathcal{R}_T^{\mathcal{E}}$ test								$\mathcal{A}_T^{\mathcal{E}}$ test							
	$\underline{\mu}$	.05	.15	.25	.15	.15	.15	.15	.05	.15	.25	.15	.15	.15	.15
	h	1	1	1	4	8	1	1	1	1	1	4	8	1	1
T	k_2	1	1	1	1	1	3	5	1	1	1	1	1	3	5
50		.093	.080	.074	.085	.083	.065	.036	.067	.067	.064	.056	.038	.057	.046
100		.098	.067	.061	.070	.078	.069	.056	.058	.058	.057	.054	.051	.056	.053
200		.070	.063	.056	.070	.065	.061	.056	.054	.054	.051	.059	.051	.053	.054
500		.058	.051	.053	.066	.062	.055	.052	.052	.052	.053	.055	.059	.047	.048

Notes to Table 4. h is the forecast horizon, $k_2 + 1$ is the number of regressors in the nesting forecasting model. The nominal significance level is 0.05. The number of Monte Carlo replications is 5,000 for $h = 1$ and 500 for $h > 1$. When the parametric bootstrap critical values are used with the number of bootstrap replications set to 199.

Table 5. Size of Non-Nested Model Comparison Tests — DGP2

	$\mathcal{R}_T$ test								$\mathcal{A}_T$ test						
	$\underline{\mu}$	.05	.15	.25	.15	.15	.15	.15	.05	.15	.25	.15	.15	.15	.15
	h	1	1	1	4	8	1	1	1	1	1	4	8	1	1
T	k	2	2	2	2	2	4	6	2	2	2	2	2	4	6
50		.010	.017	.019	.000	.000	.071	.375	.021	.029	.031	.000	.000	.038	.100
100		.018	.024	.023	.000	.000	.058	.278	.036	.039	.040	.003	.000	.046	.084
200		.023	.029	.031	.004	.000	.049	.127	.040	.041	.040	.013	.001	.045	.060
500		.031	.036	.036	.024	.005	.040	.064	.043	.042	.044	.033	.004	.046	.055

Notes to Table 5. We consider the two-sided version of our tests $\mathcal{R}_T$ and $\mathcal{A}_T$. h is the forecast horizon, k is the number of regressors in the larger forecasting model. The nominal significance level is 0.05. The number of Monte Carlo replications is 5,000.

Table 6. Size of Regression-Based Tests of Predictive Ability — DGP3

	$\mathcal{R}_T^{\mathcal{W}}$ test								$\mathcal{A}_T^{\mathcal{W}}$ test						
	$\underline{\mu}$	.05	.15	.25	.15	.15	.15	.15	.05	.15	.25	.15	.15	.15	.15
	h	1	1	1	4	8	1	1	1	1	1	4	8	1	1
T	p	1	1	1	1	1	3	5	1	1	1	1	1	3	5
50		.149	.027	.024	.048	.064	.022	.010	.044	.038	.040	.042	.064	.024	.014
100		.027	.031	.033	.030	.040	.046	.016	.042	.046	.047	.036	.050	.038	.016
200		.034	.037	.038	.058	.042	.036	.056	.040	.043	.044	.064	.054	.040	.048
500		.041	.039	.040	.052	.040	.036	.070	.045	.047	.046	.050	.050	.040	.046

Notes to Table 6. h is the forecast horizon, p is the number of restrictions being tested. The nominal significance level is 0.05. The number of Monte Carlo replications is 5,000 for $h = p = 1$ and 500 for $h > 1$ or $p > 1$. When the parametric bootstrap critical values are used with the number of bootstrap replications set to 199.

Table 7. Size of Fixed Window Tests – DGP 4

	GW_T Test		CW_T Test	
P	$\underline{\mathrm{R}}$=20	$\underline{\mathrm{R}}$=30	$\underline{\mathrm{R}}$=20	$\underline{\mathrm{R}}$=30
100	0.0824	0.1140	0.0546	0.0652
200	0.0676	0.0936	0.0460	0.0444
500	0.0362	0.0638	0.0416	0.0480

Notes to Table 7. The table reports empirical rejection frequencies of the GW_T test, eq. (23), implemented with $\underline{R}$=20 or 30, and $\overline{R}$=$\underline{R}$+5 and of the CW_T test, eq. (22). The nominal significance level is 0.05. The number of Monte Carlo replications is 5,000.

Table 8. Rejection Frequencies of Nested Model Comparison Tests — DGP1

One-Time Break Alternative

	Tests Based on Single R						Data Mining		Proposed Tests	
τ	10	20	40	80	120	160	all R	five R	$\mathcal{R}_T^{\mathcal{E}}$ Test	$\mathcal{A}_T^{\mathcal{E}}$ Test
0	.071	.063	.058	.052	.055	.057	.199	.145	.063	.054
40	.467	.445	.107	.085	.077	.096	.493	.502	.275	.075
80	.860	.925	.902	.232	.207	.232	.975	.959	.935	.647
120	.978	.993	.995	.975	.332	.331	1.000	.999	.998	.985
160	.997	1.000	1.000	1.000	0.980	0.400	1.000	1.000	1.000	1.000

Table 9. Rejection Frequencies of Non-nested Model Comparison Tests — DGP2

One-Time Break Alternative

	Tests Based on Single R						Data Mining		Proposed Tests	
τ	10	20	40	80	120	160	all R	five R	$\mathcal{R}_T$ Test	$\mathcal{A}_T$ Test
0	.051	.045	.044	.043	.041	.042	.175	.129	.029	.041
40	.111	.073	.043	.039	.041	.041	.187	.161	.031	.034
80	.380	.332	.155	.045	.038	.038	.247	.413	.080	.025
120	.695	.686	.518	.117	.048	.042	.562	.753	.304	.050
160	.890	.903	.832	.523	.138	.058	.843	.931	.654	.394

Table 10. Rejection Frequencies of Regression-Based Tests of Predictive Ability — DGP3

One-Time Break Alternative

	Tests Based on Single R						Data Mining		Proposed Tests	
τ	10	20	40	80	120	160	all R	five R	$\mathcal{R}_T^{\mathcal{W}}$	$\mathcal{A}_T^{\mathcal{W}}$
0	.026	.037	.046	.048	.053	.051	.253	.136	.037	.047
40	.035	.040	.034	.037	.042	.035	.198	.112	.022	.038
80	.146	.159	.089	.020	.018	.014	.211	.193	.022	.038
120	.431	.494	.352	.073	.006	.006	.513	.516	.108	.059
160	.842	.903	.849	.495	.089	.003	.932	.925	.493	.410

Notes to Tables 8-10. τ is the break date with $\tau = 0$ corresponding to the null hypothesis. We set $h = 1$,$\underline{\mu}$= 0.15, $\overline{\mu}$= 0.85, $T = 200$ and $p = 0$ in Table 8 and $p = 1$ in Tables 9-10. The five values of R used in the last column are $R = 20, 40, 80, 120, 160$. The number of Monte Carlo replications is set to 5,000.

Table 11. Data Mining – Asymptotic Approximation Results

$\underline{\mu}$	DMW_T		$\mathcal{W}_T^{(r)}$				ENCNEW_T			
		$p=$	1	2	3	4	1	2	3	4
0.15	0.2604		0.2604	0.2712	0.2750	0.2784	0.1023	0.1251	0.1347	0.1305
0.20	0.0963		0.2296	0.2391	0.2412	0.2462	0.1161	0.1264	0.1224	0.2017
0.25	0.0928		0.2017	0.2102	0.2112	0.2166	0.0903	0.1124	0.1215	0.1178
0.30	0.1761		0.1761	0.1842	0.1838	0.1881	0.0903	0.1087	0.1170	0.1148
0.35	0.1513		0.1513	0.1584	0.1581	0.1606	0.0853	0.0996	0.1075	0.1066

Notes to Table 11: The table shows asymptotic rejections of nominal 5% tests for non-nested models (DMW_T), forecast optimality ($W_T^{(r)}$) and nested models ($ENCNEW_T$) repeated over sequences of windows sizes equal to $[\underline{\mu}T]$, $[\underline{\mu}T+1], ..., [(1-\underline{\mu})T]$. Asymptotic approximations to the tests statistics are based on Brownian motion approximation with $T = 10,000$. The number of Monte Carlo replications is 5,000.

Table 12. Empirical Results

	$\mathcal{R}_T$ Test		$\mathcal{A}_T$ Test		"Test Based on Single R"	
	UIRP	Taylor	UIRP	Taylor	UIRP	Taylor
Japan	10.43**	7.30**	-3.20	-4.59	-5.88	2.55
Canada	73.06**	44.44**	7.13**	15.75**	15.62**	30.07**
Switzerland	16.59**	- -	-1.00	- -	-15.76	- -
U.K.	9.06**	22.26**	-11.65	-1.68	-20.58	6.88**
France	-1.10	-0.01	-12.33	-9.57	-13.49	-14.29
Germany	3.83	0.87	-11.91	-15.54	-17.28	-21.30
Italy	24.99**	27.40**	-2.07	-5.33	12.31**	-6.88
Sweden	57.79**	42.26**	-2.38	5.58**	-22.28	-12.70
The Netherlands	- -	7.59**	- -	-2.70	- -	1.35
Portugal	- -	109.37**	- -	24.30**	- -	-10.43

Notes to Table 12. Two asterisks denote significance at the 5% level, and one asterisk denotes significance at the 10% level. For the R_T and A_T tests we used $\underline{\mu}= 0.15$ (the value of $\underline{R}$ will depend on the sample size, which is different for each country, and it is shown in Figures 1 and 2). For the "Test Based on Single R", we implemented Clark and McCracken's (2001) test using $R = 120$; its one-sided critical values at the 5% and 10% significance levels are 3.72 and 2.65.

Figure 1

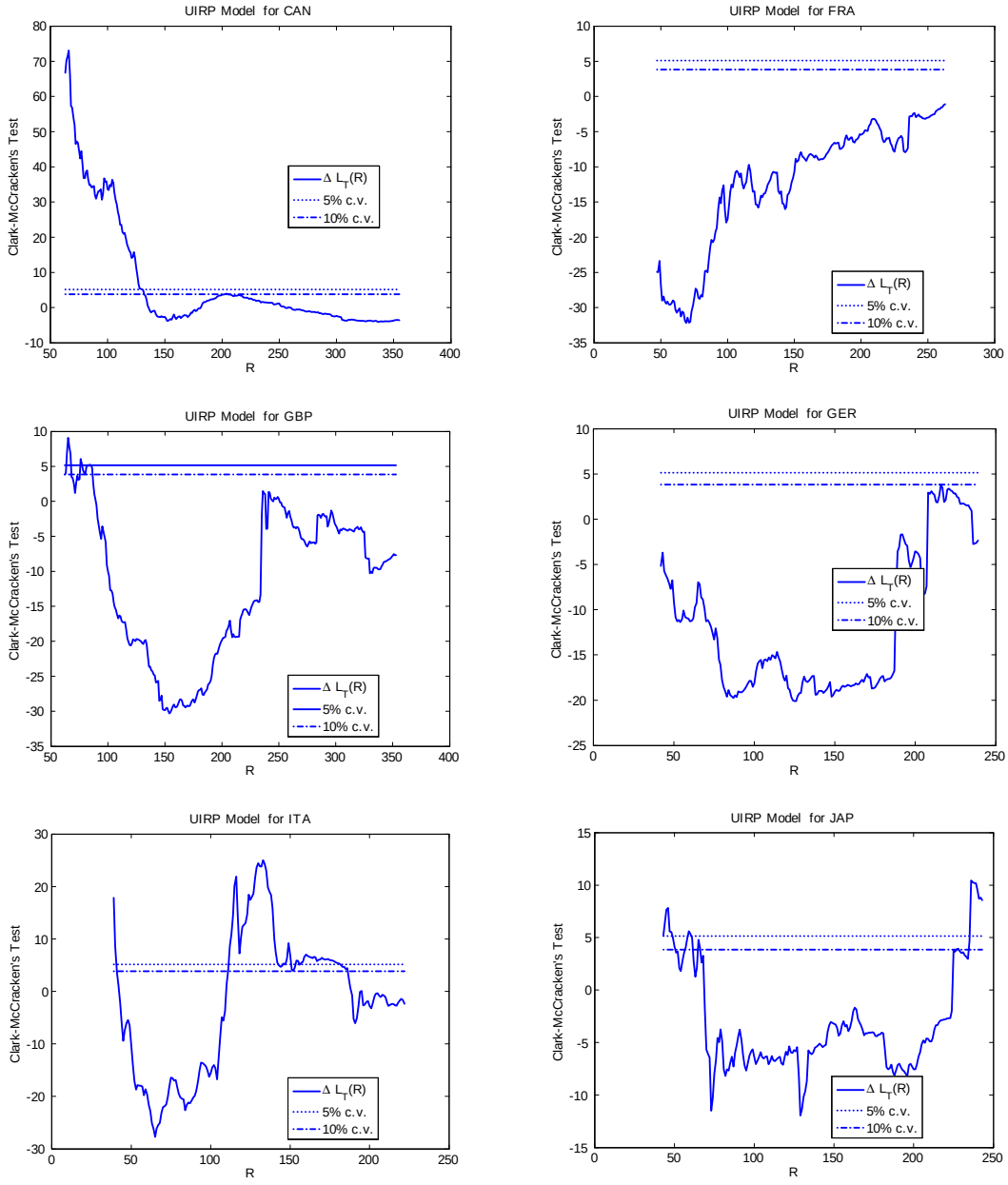

Figure 1 plots the estimated Clark and McCracken (2001) ENCNEW test statistic for comparing the UIRP model with the random walk for the window sizes we consider (reported on the x-axis), together with 5% and 10% critical values of the $\mathcal{R}_T^{\mathcal{E}}$ test statistic. The test rejects when the largest value of the Clark and McCracken's (2001) test is above the critical value line. Countries are Canada (CAN), France (FRA), United Kingdom (GBP), Germany (GER), Italy (ITA), Japan (JAP).