WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 11-31
OUT-OF-SAMPLE FORECAST TESTS ROBUST TO THE CHOICE OF WINDOW SIZE

Barbara Rossi
Duke University and
Visiting Scholar, Federal Reserve Bank of Philadelphia

Atsushi Inoue
North Carolina State University

August 5, 2011

RESEARCH DEPARTMENT, FEDERAL RESERVE BANK OF PHILADELPHIA
Ten Independence Mall, Philadelphia, PA 19106-1574 • www.philadelphiafed.org/research-and-data/

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

Barbara Rossi and Atsushi Inoue

Duke University and NC State

August 5, 2011

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. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at www.philadelphia.org/research-and-data/publications/working-papers/

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, Meese and Rogoff (1983a) use a window of 93 observations, Chinn (1991) a window size equal to 45, Qi and Wu (2003) use a window of 216 observations, Cheung et al. (2005) consider windows of 42 and 59 observations, van Dijk and Franses' (2005) window size is 128, Clark and West's (2006, 2007) window is 120 observations, Gourinchas and Rey (2007) consider a window of 104 observations, and Molodtsova and Papell (2009) consider a window size of 120 observations. 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 oversized. 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 $S_T(.)$ we focus on satisfies*

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

where $\Rightarrow$ 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.

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 sub-sections, 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, a sketch of the proof that the test statistics satisfy condition (2) is provided in Appendix A. 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 $[\underline{\mu}, \overline{\mu}]$.

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: $E\left[\Delta L_T^* (R)\right]$. 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 (2006) 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.

Appendix A 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 (which rules out unit roots and structural breaks), and the forecast errors (which can be either one period or multi-period) should be evaluated using continuously differentiable loss functions, which include Mean Square Forecast Errors (MSFE) as a special case.

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 $\hat{\sigma}_R^2$ is a consistent estimator of σ^2. Reject the null hypothesis $H_0:\lim_{T\to\infty}E\left[\Delta L_T^*(R)\right]=0$ for all R in favor of the alternative $H_A:\lim_{T\to\infty}E\left[\Delta L_T^*(R)\right]\neq 0$ for some R at the significance level α when $\mathcal{R}_T>k_\alpha^{\mathcal{R}}$ or when $\mathcal{A}_T>k_\alpha^{\mathcal{A}}$, where the critical values $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$ are reported in Table 1, Panel A, for various values of $\underline{\mu}$.

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: $\mathcal{R}_T=\sup_{R\in\left[\underline{R},\ldots\overline{R}\right]}\Delta L_T(R)$, $\mathcal{A}_T=\frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}}\Delta L_T(R)$. The tests reject the null hypothesis $H_0:\lim_{T\to\infty}E\left[\Delta L_T^*(R)\right]=0$ for all R in favor of the alternative $H_A:\lim_{T\to\infty}E\left[\Delta L_T^*(R)\right]<0$ for some R at the significance level α when $\mathcal{R}_T>k_\alpha^{\mathcal{R}}$ or when $\mathcal{A}_T>k_\alpha^{\mathcal{A}}$, where the critical values $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$ are reported in Table 1, Panel B, for various values of $\underline{\mu}$.

INSERT TABLE 1 HERE

Finally, it is useful to remind readers that, as discussed in Clark and McCracken (2001), (5) is not necessarily asymptotically normal even when the models are not nested. For example, when $y_{t+1} = \alpha_0 + \alpha_1 + u_{t+1}$ 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 regressors $x_{1,t}$ and the second ("large") model uses 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 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 various values of $\underline{\mu}$ are reported in Table 2(a) for the rolling window estimation scheme, and in Table 2(b) for the recursive window estimation scheme.

INSERT TABLE 2(a,b) HERE

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 2 (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,

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

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 1 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

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

and

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

where

$$\mathcal{L}_T\left(R\right) \equiv P^{-1/2}\sum_{t=R}^{T}\mathcal{L}_{t+h}(\widehat{\theta}_{t,R}),$$

and $\widehat{\Omega}_R$ is a consistent estimator of Ω. Reject the null hypothesis $H_0 : \lim_{T\to\infty} E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0$ 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 various values of $\underline{\mu}$ are reported in Table 3.

INSERT TABLE 3 HERE

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 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:

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

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 $H_0 : \lim_{T\to\infty} E[\widehat{\alpha}(R)] = 0$ 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 various values of $\underline{\mu}$ and 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 testsconsidered 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 (2005, 2007) and Giacomini and White (2005), which rely on a constant (fixed) window size. Propositions 3 and 4 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 (2005, 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 (2005, 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. 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 $\hat{\beta}_{2t}(R) = (\sum_{s=t-R+1}^{t} x_{1,s} x_{1,s}')^{-1} \sum_{s=t-R+1}^{t} x_{1,s} y_{s+h}$ and let $\hat{e}_{1,t+h}(R)$ and $\hat{e}_{2,t+h}(R)$ denote the corresponding models' one-step-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 $MSPE\text{-}adjusted$ of Clark and West (2007) can be written as:

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

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 in general. 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

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

where $e_{t+h} = e_{1,t+h} = e_{2,t+h}$ under the null hypothesis, and the second equality follows from the strict exogeneity of $x_{2,t}$. 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 under the null hypothesis.

Proposition 3 (Out-of-Sample Robust Test with Fixed Window Size I.) *Suppose that: (a) $E(e_{2t}|x_{2s}) = 0$ 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:

$$CW_T \equiv \xi_R'\widehat{\Omega}^{-1}\xi_R \xrightarrow{d} \chi_r^2, \tag{20}$$

where $\widehat{\Omega}$ is a consistent estimate of Ω. The null hypothesis is rejected at the significance level α for any R when $CW_T > \chi_{r;\alpha}^2$, where $\chi_{r;\alpha}^2$ 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). There are several cases of model comparisons in practice where Assumption (a) is satisfied, for example a sufficient condition is that the regressors $x_{1,t}$ and z_t be strictly exogeneous.

We also consider the Giacomini and White's (2005) framework. Proposition 4 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 4 (Out-of-sample Robust Tests with Fixed Window Size II.) *Suppose the Assumptions of Theorem 4 in Giacomini and White (2005) hold. Let*

$$GW_T = \inf_{R \in \{\underline{R}, \ldots \overline{R}\}} |\Delta L_T(R)|, \tag{21}$$

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 $H_0 : \lim_{T\to\infty} E\left[\Delta L_T\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)\right] = 0$ for one window size,

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

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 $inf(.)$ over the sequence of out-of-sample tests rather than the $sup(.)$. 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 $inf(.)$. The 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. 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.

The proof of Proposition 4 is provided in Appendix A. 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 $GW_T = \inf_{R=\underline{R},\ldots\overline{R}} \Delta L_T(R)$, 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.

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

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

where $y_0 = x_0 = 0$, $z_0 = 0_{p\times 1}$, and $[u_{y,t+1}\ u_{x,t+1}\ u'_{z,t+1}]' \overset{iid}{\sim} N(0_{(p+2)\times 1}, I_{p+2})$. We compare the following two nested models' forecasts for y_{t+h}:

$$\begin{aligned} \text{Model 1 forecast} &: \widehat{\theta}_{1,t} y_t \\ \text{Model 2 forecast} &: \widehat{\gamma}_{1,t} y_t + \widehat{\gamma}'_{2,t} x_t + \widehat{\gamma}'_{3,t} z_t, \end{aligned} \tag{22}$$

and both models' parameters are estimated by OLS in rolling windows of size R. Under the null hypothesis $d_t = 0$ for all t and we consider $h = 1, 4, 8$, $p = 0, 3, 5$ and $T = 50, 100, 200, 500$. Under the no-break alternative hypothesis $d_t = 0.1$ or $d_t = 0.2$ ($h = 1$, $p = 0$ and $T = 200$). Under the one-time-break alternative hypothesis, $d_t = 0.5 \cdot I(t \le \tau)$ for $\tau \in \{40, 80, 120, 160\}$, ($h = 1$, $p = 0$ and $T = 200$).

For the non-nested models' comparison, we consider a modification of DGP1 (labeled "DGP2"):

$$\begin{pmatrix} y_{t+1} \\ x_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} 0.3 & d_t & 0_{1\times p} \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5I_p \end{pmatrix} \begin{pmatrix} y_t \\ x_t \\ z_t \end{pmatrix} + \begin{pmatrix} u_{y,t+1} \\ u_{x,t+1} \\ u_{z,t+1} \end{pmatrix}, \quad t = 1, ..., T-1,$$

where $y_0 = x_0 = 0$, $z_0 = 0_{p\times 1}$, and $[u_{y,t+1}\ u_{x,t+1}\ u'_{z,t+1}]' \overset{iid}{\sim} N(0_{(p+2)\times 1}, I_{p+2})$. We compare the following two non-nested models' forecasts for y_{t+h}:

$$\begin{aligned} \text{Model 1 forecast} &: \widehat{\theta}_1 y_t + \widehat{\theta}_2 x_t \\ \text{Model 2 forecast} &: \widehat{\gamma}_1 y_t + \widehat{\gamma}'_2 z_t, \end{aligned} \tag{23}$$

and both models' parameters are estimated by OLS in rolling windows of size R. Under the null hypothesis $d_t = 0.5$ for all t and we consider $h = 1, 4, 8$, $p = 1, 3, 5$ and $T = 50, 100, 200, 500$. Under the no-break alternative hypothesis $d_t = 0.5$ or $d_t = 1$ ($h = 1$, $p = 1$ and $T = 200$). Under the one-time-break alternative hypothesis, $d_t = 0.5 \cdot I(t \leq \tau) + 0.5$ for $\tau \in \{40, 80, 120, 160\}$, ($h = 1$, $p = 1$ and $T = 200$).

"DGP3" is designed for regression-based tests and is a modification of West and McCracken's (1998). Let

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

where $\varepsilon_{t+1} \overset{iid}{\sim} N(0_{p\times 1}, I_p)$. We focus on tests of forecast unbiasedness discussed in Definition 2. 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 $\delta_t = 0.5$ or $\delta_t = 1$ under the no-break alternative and $\delta_t = 0.5 \cdot I(t \leq \tau)$ 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"):

$$y_{t+1} = \delta_t x_t + \varepsilon_{t+1}, \quad t = 1, ..., T,$$

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:

$$\begin{aligned} \text{Model 1 forecast} &: 0 \\ \text{Model 2 forecast} &: \widehat{\delta}_t x_t. \end{aligned}$$

To ensure that the null hypothesis in Proposition 4 holds for one of the window sizes, $\underline{R}$, we let $\delta_t = (\underline{R} - 2)^{-1/2}$. The number of Monte Carlo replications is 5,000. To ensure that the null hypothesis in Proposition 3 holds, we let $\delta_t = 0$.

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 10%, 5% and 1% nominal levels. 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 $\overline{R} = \underline{R} + 5$. 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. In each table, panel A reports results for the baseline case, and panels B and C report results for the multiple forecast horizon case and for the multiple regressor case, respectively. For the multiple horizon case (panel B), 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 (2011, Section 4).

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, and the $\mathcal{R}_T$ test tends to reject too often when there are many regressors ($p = 5$). 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.

INSERT TABLES 4, 5 and 6

Table 7 reports empirical rejection frequencies for DGP4. The left panel shows results for the GW_T test, eq. (21), 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 3.

INSERT TABLE 7

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, the non-nested model comparison test of Diebold and Mariano (1995), West (1996) and McCracken (2000) under DGP2, and West and McCracken's (1998) regression-based test of predictive ability under DGP3, 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 four values, $R \in \{40, 80, 120, 160\}$. 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% 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. Panel A of each table shows that the conventional tests and proposed tests have power against the standard no-break alternative hypothesis. 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 $\underline{\mu}$.

Panel B of each table demonstrates 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. Our proposed tests tends to be more powerful than those based on the ad-hoc 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.

INSERT TABLES 8, 9 and 10

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 our test statistics. Table 11 reports the results, which demonstrate that the over-rejections of traditional tests when researchers data snoop over window sizes persist asymptotically.

INSERT TABLE 11

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,

$$s_{t+1} - s_t = \alpha + \beta z_t + \varepsilon_{t+1} \tag{24}$$

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,

$$f_t^{UIRP} = (i_t - i_t^*) + s_t, \tag{25}$$

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:

$$f_t^{TAYLOR} = \delta(\pi_t - \pi_t^*) + \gamma(y_t^{gap} - y_t^{gap*}) + s_t. \tag{26}$$

The benchmark model, against which the forecasts of both models (25) and (26) 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 are shown in Table 12 and Figures 1 and 2. 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. Figures 1 and 2 report the estimated Clark and McCracken's (2001) test statistic for the window sizes we consider. Note that the $\mathcal{R}_T^{\varepsilon}$ 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.

INSERT TABLE 12 AND FIGURES 1 AND 2

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

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

[3] 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.

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

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

[6] Clark, T.E. and M.W. McCracken (2000), "Not-for-Publication Appendix to Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *mimeo,* Kansas City Fed.

[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 (2011), "Advances in Forecast Evaluation," in: G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 2, Elsevier, forthcoming.

[13] 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.

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

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

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

[17] 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.

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

[19] 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).

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

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

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

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

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

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

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[44] Stock, J.H. and M.W. Watson (2003b), *Introduction to Econometrics*, Addison Wesley.

[45] Van Dijk, D. and P.H. Franses (2003), "Selecting a Nonlinear Time Series Model Using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics* 65, 727-744.

[46] West, K.D. (1994), "Asymptotic Inference About Predictive Ability: Additional Appendix," mimeo, University of Wisconsin.

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

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

7 Appendix A. Proofs

Verification of the High-Level Assumptions in Proposition 1

In this appendix we verify that the tests of predictive ability we focus on satisfy (2). Then the results (3) and (4) follow by the Continuos Mapping Theorem. We follow the proofs of West (1996), Clark and McCracken (2001) and West and McCracken (1998) very closely and extend their results to weak convergence in the space of functions on $[\underline{\mu}, \overline{\mu}]$.

(a) The tests of West (1996) and McCracken (2000): For the out-of-sample tests of non-nested model comparisons considered in Section 2.1, we adopt the same notation as in West (1996) and assume that Assumptions 1–4 in West (1995) hold with Assumption 4 holding for all values of $R = [\underline{\mu}T], [\underline{\mu}T] + 1, ..., [\overline{\mu}T]$ (Assumptions 1–4 of McCracken, 2000 are identical to Assumptions 1–4 in West, 1996). It is relatively straightforward to show that Lemmas A1–A6 of West (1996) holds under these assumptions with the $\sup_t$ replaced by $\sup_{\underline{R} \leq R \overline{R}} \sup_{R \leq t \leq T} \equiv \sup_{[\underline{\mu}T] \leq t \leq]}$. Thus the key is to prove our version of his Lemma 4.1 and it suffices to verify that the Assumptions C1, C2 and A5 in Corollary 3.1 in Wooldridge and White (1988) are satisfied by $P^{1/2}(\overline{f} - Ef_t)$.

Assumption C1 is a nominal assumption and is trivially satisfied for the test statistics of equal predictive ability of non-nested models estimated with either rolling, recursive and split-sample estimation techniques.

Let $H_t = (1/t)\sum_{s=1}^{t} h_s(\beta^*)$ in the recursive scheme and $H_t = (1/R)\sum_{s=t-R+1}^{t} h_s(\beta^*)$ in the rolling scheme. We need to show that the partial sum $\sum[(f_{t+\tau} - E(f_{t+\tau}))' \, H_t']$ satisfies Assumption C2 of Wooldridge and White (1988). We focus on H_t because we have already assumed that f_t satisfies these assumptions. In the recursive scheme, note that

$$\sum_{t=R}^{P} H_t \;=\; a_{R,0}(h_1 + \cdots + h_R) + a_{R,1}h_{R+1} + \cdots a_{R,P-1}h_{R+P-1},$$

where $a_{R,j} = \sum_{k=j}^{P-1} 1/(R+k)$ for $j = 0, 1, ..., P-1$ (see West, 1996, p. 1081). This in turn can be written as

$$\sum_{t=R}^{P} H_t \;=\; \sum_{t=1}^{T} b_t h_t$$

where $b_t = a_{R,0}$ for $t = 1, ..., R$ and $b_t = a_{R,t-R+1}$ for $t = R+1, ..., T$. In the rolling scheme, the sum can be written as

$$\sum_{t=R}^{P} H_t \;=\; \sum_{t=1}^{T} b_t h_t$$

where $b_t = \min(t, R, T-t)/R$. Thus we show that Assumption C2 is satisfied for $b_t h_t$. In the both schemes, b_t is bounded. Thus Assumption C2(i) is satisfied. Because $\{b_t\}$ is a sequence of deterministic bounded constants and h_t is strong mixing, $b_t h_t$ is near epoch dependent as required by Assumptions C2(ii)(iii). Assumption C2(iv) is also satisfied because of the assumptions on h_t and the boundedness of b_t.

Finally, Assumption A5 is satisfied since

$$Var\left(T^{-1/2}\sum_{t=R}^{[sT]} H(t)\right) = Var\left(s^{1/2}\left[sT\right]^{-1/2}\sum_{t=R}^{[sT]} H(t)\right) \to s \cdot \lambda_{hh} S_{hh}$$

using McCracken's (2000) notation.

Therefore Theorem 3.1 of Wooldridge and White (1989) completes the proof.

(b) The ENCNEW test of Clark and McCracken (2001): For the out-of-sample tests of nested model comparisons considered in Section 2.2, we assume that Assumptions 1, 2, 3 and 4 in Clark and McCracken (2001) hold with Assumption 4 holding for all values of $R = [\underline{\mu}T]$, $[\underline{\mu}T]+1, ..., [\overline{\mu}T]$. To prove our version of their Theorem 3.3(a), we need to show that Lemma 1, Lemma 8, Lemma 9(a)(b) and 10(b) of Clark and McCracken (2000b) hold uniformly in $\mu \in [\underline{\mu}, \overline{\mu}]$ rather than pointwise in μ, where our μ is their $\lambda = (1+\pi)^{-1}$. Note that T/R in the proof of Lemma 1(a) and that in the first inequality on page 6 of Clark and McCracken (2000b) are bounded uniformly in μ and that

$$\sup_{R \le t \le T} |T^{-1/2}\sum_{j=1}^{t}[Q_t - E(Q_t)] \le \sup_{1 \le t \le T} |T^{-1/2}\sum_{j=1}^{t}[Q_t - E(Q_t)] = O_p(1)$$

where Q_t is a generic random variable that satisfies the Functional Central Limit Theorem and $O_p(1)$ is uniform in $\mu \in [\underline{\mu}, \overline{\mu}]$. Using these two results, we can prove that Lemmas 1(a) and (b) hold uniformly in $\mu \in [\underline{\mu}, \overline{\mu}]$. The uniform version of Lemma 1(c) follows from the uniform version of Lemmas 1(a) and 1(b). The proofs of Lemmas A2, A3(b), A5 and A6 are based on the Function Central Limit Theorem of Davidson (1994, Corollary 29.19) and the convergence of martingale approximations (Hansen, 1992, Theorem 3.1) both of which holds uniformly in μ. Therefore given the proofs of Clark and McCracken (2000b) it is straightforward to extend their proofs to show that Lemmas A2, A5 and A6 holds uniformly in $\mu \in [\underline{\mu}, \overline{\mu}]$. Note that Lemma A4 holds. Therefore Lemma A10 holds uniformly in $\mu \in [\underline{\mu}, \overline{\mu}]$. The uniform version of Theorem 3.3 follows from the uniform version of Lemmas A6 and A10.

(c) The test of West and McCracken (1998): For the tests of regression based predictive

ability considered in Section 2.3, we assume that Assumptions 1–5 in West and McCracken (1998) hold with w_t in Assumption 4 replaced by

$$w_t \;=\; [v'_{t\beta}\ \mathrm{vec}(g_{t\beta})'\ v_t\ g'_t\ h'_t\ vech((g_t g'_t)']'$$

and Assumption 5 holding for all π such that $(1-\overline{\mu})/\overline{\mu} \leq \pi \leq (1-\underline{\mu})/\underline{\mu}$ and all values of $R = [\underline{\mu}T],\ [\underline{\mu}T]+1, ..., [\overline{\mu}T]$. We need to add $vech((g_t g'_t)$ in the definition of w_t to extend their Lemma 4.3. Then the proof follows by applying similar arguments to those in (a).

(i) Lemmas A1, A2 and A3 in West and McCracken (1998) follow directly from our discussion in (a).

(ii) Uniform convergence of HAC covariance estimators in both West (1995) and West and McCracken (1998) can be proved as follows. Lemmas A3 and A4 in West and McCracken (1998) can be strengthened to ensure uniform convergence of HAC covariance estimators in the following way. By the uniform convergence of the parameter estimates (shown in (a)), the fact that $\mu \in \left[\underline{\mu}, \overline{\mu}\right]$ which is bounded, the assumption that the variance is non-singular for each μ, using Lemma A3 in Andrews (1993) and the discussion on p. 835 in Andrews (1993) we conclude that Lemma A3 in West and McCracken (1998) holds. Similarly, Lemma A4 in West and McCracken (1998) holds because of the parameter estimate is uniformly consistent, and the fact that f_t and its derivatives are uniformly consistent. The test statistic in Clark and West (2007) focuses on the product of the forecast error of the small model times the difference in the forecasts of the small model and the large model. Thus, the test statistic has the same structure as the tests of forecast rationality considered in West and McCracken (1998).

(iii) Lemma 4.1 can be strengthened as follows. Lemma 4.1(b) is strengthened as $P^{-1/2}\sum_{t=R}^{[Ts]} g_{t+1}v_{t+h}$ $\Rightarrow S_{ff}^{1/2}B(s)$ and (a) and (c) hold uniformly in R :

$$\begin{aligned}
\sup_R \left|P^{-1/2}\sum_{t=R}^{[Ts]}\widehat{g}_{t+1}\widehat{v}_{t,t+\tau} - P^{-1/2}\sum_{t=R}^{[Ts]} g_{t+1}v_{t,t+\tau} - FB\left[P^{-1/2}\sum H(t)\right]\right| &= o_p(1)\\
\sup_R \left|E\sum H(t)\,H(t)'] - \lambda_{hh}S_{hh}\right| &= o_p(1)\\
\sup_R \left|E\left(P^{-1}\sum g_{t+1}v_{t+\tau}\sum H(t)'\right)] - \lambda_{fh}S_{fh}\right| &= o_p(1).
\end{aligned}$$

We extend Lemma 4.2 by applying FCLT in Corollary 3.1 Wooldridge and White (1988) similarly to the proof of (a).

(iv) To prove Lemma 4.3, as in West and McCracken, p. 836) we focus on the case in which g_t and β_t are scalars. By our version of Lemma A1, $\sup_R \sup_t |\widehat{\beta}_t - \beta^*| = o_p(1)\sup_R P^{-1}\sum g^2_{t+1,\beta\beta}\left(\widetilde{\beta}_t\right) =$

$O_p(1)$ by the fact that $\sup_R P^{-1}\sum g_{t+1,\beta\beta}^2\left(\widetilde{\beta}_t\right) \leq (T-\overline{\mu}T-h)^{-1}\left|P^{-1}\sum_{t=\underline{\mu}T}^{T} g_{t+1,\beta\beta}^2\left(\widetilde{\beta}_t\right)\right| = O_p(1)$ by assumption 3 (which is uniform because it does not depend on R), and the remaining terms are $O_p(1)$ uniformly in R by our version of Lemma A1 in (i) and Markov's inequality.

(v) Our version of Theorems 4.1, 4.2 and 5.1 follow from these lemmas and replacing pointwise convergence by uniform convergence, as we did in (a). Theorem 7.1 let $\widehat{\alpha}(R) \equiv \left(\sum_{t=R}^{T}\widehat{g}_{t+1}^2\right)^{-1}\left(\sum_{t=R}^{T}\widehat{g}_{t+1}\widehat{v}_{t+1}\right)$ Clark and McCracken's results can be strengthened as $P^{1/2}\widehat{\alpha}(R) \Rightarrow V^{1/2}B(\mu)$ by using the same arguments as in the proof of Lemma 4.2.

Proof of Proposition 4 Following the assumptions in Theorem 4 of Giacomini and White (2005), we have

$$\Delta L_T(R) \xrightarrow{d} N(0,1) \tag{27}$$

for every $R = \underline{R}, \underline{R}+1, ..., \overline{R}$, where $\underline{R}$ and $\overline{R}$ are two constants that are fixed relative to the sample size. Since the asymptotic distribution in eq. (27) does not depend on the choice of R, and the set of values that R can take is fixed, it follows that

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

8 Appendix B. Variance Estimation and Asymptotic Distributions.

(a) For the out-of-sample tests of non-nested model comparisons considered in Section 2.1, a consistent estimator for the variance that takes into account parameter estimation error is as follows (McCracken, 2000): $\widehat{\sigma}_R^2 = \widehat{\sigma}_{ff}^2 + \lambda_{fh}\left(\widehat{F}\widehat{B}\widehat{S}'_{fh} + \widehat{S}_{fh}\widehat{B}'\widehat{F}'\right)a + \lambda_{hh}\widehat{F}\widehat{B}\widehat{S}_{hh}\widehat{B}'\widehat{F}'$, where λ_{fh} and λ_{hh} are such that:

Scheme:	λ_{fh}	λ_{hh}
Recursive	$1-\pi^{-1}\ln(1+\pi)$	$2\left[1-\pi^{-1}\ln(1+\pi)\right]$
Rolling, $\pi \leq 1$	$\pi/2$	$\pi-\pi^2/3$
Rolling, $1<\pi<\infty$	$1-(2\pi)^{-1}$	$1-(2\pi)^{-1}$
Fixed	0	π,

(28)

$\widehat{\sigma}_{ff,R}^2$, $\widehat{S}_{fh}$, S_{hh} are the HAC variance estimates of f_{t+h}, $f_{t+h}h_t$, h_t respectively, $\widehat{F} = T^{-1}\sum_{t=R}^{[sT]}\left(\partial f_{t+h}\left(\widehat{\beta}_{t,R}\right)/\partial\beta\right)$, and $\widehat{B} = B_T$, where B_T is such that $\widehat{\beta}_{t,R} - \beta^* = B_t H_t$. Note that the critical values for significance level α are, respectively, $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$, where $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$ solve $P\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]}(1-\mu)^{-1/2}\left|B(1)-B(\mu)\right| > k_\alpha^{\mathcal{R}}\} = \alpha$ and $P\{\int_{\underline{\mu}}^{\overline{\mu}}(1-\mu)^{-1/2}$ $\left|\mathcal{B}(1)-\mathcal{B}(\mu)\right| d\mu > k_\alpha^{\mathcal{A}}\} = \alpha$, and are computed using Monte Carlo simulation methods.

(b) For the out-of-sample tests of nested model comparisons considered in Section 2.2, it follows from Proposition 1 that $R_T^{\mathcal{E}} \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]}\mu^{-1}\int_\mu^1\left[\mathcal{B}_k(s)-\mathcal{B}_k(s-\mu)\right]' dB_k(s)$, and $A_T^{\mathcal{E}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}}\left\{\{\mu^{-1}\int_\mu^1\left[\mathcal{B}_k(s)-\mathcal{B}_k(s-\mu)\right]' d\mathcal{B}_k(s)\}\right\}$, where $\underline{R} = [\underline{\mu}T]$, $\overline{R} = [\overline{\mu}T]$, $B_k(\cdot)$ is a standard k-variate Brownian motion and k is the number of parameters in the larger model in excess of the parameters in the smaller model. Note that the critical values for a significance level α are, respectively, $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$, where $k_\alpha^{\mathcal{R}}$ and $k_\alpha^{\mathcal{A}}$ solve $P\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]}$ $\mu^{-1}\int_\mu^1\left[\mathcal{B}_k(s)-\mathcal{B}_k(s-\mu)\right]' dB_k(s) > k_\alpha^{\mathcal{R}}\} = \alpha$ and $P\{\int_{\underline{\mu}}^{\overline{\mu}}\left\{\mu^{-1}\int_\mu^1\left[\mathcal{B}_k(s)-\mathcal{B}_k(s-\mu)\right]' d\mathcal{B}_k(s)\right\}$ $> k_\alpha^{\mathcal{A}}\} = \alpha$, respectively. The critical values are obtained via Monte Carlo simulation methods.

(c) For the out-of-sample tests of nested model comparisons considered in Section 2.3, a consistent estimator for the variance that takes into account parameter estimation error in the case of Mincer and Zarnowitz's (1969) regressions is as follows (West and McCracken, 1998): $\widehat{\Omega}_R = S_{ff} + \lambda_{fh}\left(FBS'_{fh} + S_{fh}B'F'\right) + \lambda_{hh}FBS_{hh}B'F'$, where λ_{fh}, λ_{hh} have been

defined in eq. (28), S_{ff}, S_{fh}, S_{hh} are HAC variance estimates of $\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})$, $\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})h_t$, h_t, $\widehat{F} = T^{-1}\sum_{t=R}^{[sT]}\left(\partial\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})/\partial\theta\right)$, and $\widehat{B} = B_T$, where B_T is such that $\widehat{\theta}_{t,R} - \theta^* = B_t H_t$.

Note that the statistics proposed in this paper build upon: $T^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})'\widehat{\Omega}_R^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})$, whereas the traditional Wald test is: $W_T(R) = P^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})'\widehat{\Omega}_R^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})$. Under the null hypothesis $H_0 : \lim_{T\to\infty} E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0$ for all R, $R_T^{\mathcal{W}} \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]}$ $(1-\mu)^{-1}\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]'\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]$, and $A_T^{\mathcal{W}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}}(1-\mu)^{-1}\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]'$ $\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]d\mu$,where $R = [\mu T]$, $\underline{R} = [\underline{\mu}T]$, $\overline{R} = [\overline{\mu}T]$, and $B_p(\cdot)$ is a standard p-dimensional Brownian motion. Critical values for a significance level α are, respectively, $k_\alpha^{\mathcal{R},\mathcal{W}}$ and $k_\alpha^{\mathcal{A},\mathcal{W}}$, where $k_\alpha^{\mathcal{R},\mathcal{W}}$ and $k_\alpha^{\mathcal{A},\mathcal{W}}$ solve $P\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]}(1-\mu)^{-1}\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]'\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]$ $> k_\alpha^{\mathcal{R},\mathcal{W}}\} = \alpha$, $P\{\int_{\underline{\mu}}^{\overline{\mu}}(1-\mu)^{-1}\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]'\left[\mathcal{B}_p(1) - \mathcal{B}_p(\mu)\right]d\mu > k_\alpha^{\mathcal{A},\mathcal{W}}\} = \alpha$.

9 Tables and Figures

Table 1. Critical Values for Non-Nested Model Comparisons

Panel A. Two-Sided Critical Values						
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
$\underline{\mu}$	10%	5%	1%	10%	5%	1%
0.15	2.4653	2.7540	3.3372	1.4624	1.7393	2.2928
0.20	2.3987	2.6979	3.2825	1.4891	1.7719	2.3450
0.25	2.3334	2.6418	3.2286	1.5129	1.8092	2.3945
0.30	2.2642	2.5777	3.1599	1.5399	1.8380	2.4334
0.35	2.1865	2.4989	3.0991	1.5647	1.8648	2.4755
Panel B. One-Sided Critical Values						
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
$\underline{\mu}$	10%	5%	1%	10%	5%	1%
0.15	2.1277	2.4589	3.1061	1.1344	1.4541	2.0732
0.20	2.0572	2.3990	3.0590	1.1589	1.4880	2.1166
0.25	1.9864	2.3329	3.0078	1.1797	1.5108	2.1670
0.30	1.9207	2.2614	2.9537	1.2014	1.5357	2.2058
0.35	1.8386	2.1868	2.8622	1.2258	1.5606	2.2409

Notes to Table 1. The critical values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normal random variates.

Table 2(a). Critical Values for Nested Model Comparisons Using ENCNEW in Rolling Regressions. Panel A. 10% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
	$\underline{\mu}$					$\underline{\mu}$				
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	3.9383	3.2651	2.7901	2.4207	2.1397	1.0606	1.0509	1.0557	1.0628	1.0779
2	5.6238	4.7021	4.0398	3.5472	3.1364	1.6027	1.6004	1.6131	1.6282	1.6529
3	6.9083	5.8076	5.0120	4.4155	3.9117	2.0365	2.0411	2.0526	2.0741	2.0992
4	7.9417	6.6788	5.7622	5.0918	4.5009	2.3769	2.3769	2.3896	2.4178	2.4400
5	8.8922	7.4685	6.4714	5.7091	5.0630	2.6650	2.6669	2.6841	2.7027	2.7506
6	9.7030	8.1372	7.0832	6.2397	5.5625	2.9012	2.9103	2.9292	2.9793	3.0271
7	10.4663	8.8057	7.6191	6.7431	6.0087	3.1514	3.1447	3.1691	3.1952	3.2424
8	11.2258	9.4397	8.1361	7.1555	6.4060	3.3606	3.3732	3.3791	3.4241	3.4861
9	11.8880	9.9882	8.6720	7.6060	6.7839	3.5543	3.5609	3.5807	3.6138	3.6682
10	12.5023	10.5105	9.1460	8.0328	7.1817	3.7282	3.7421	3.7661	3.8148	3.8799
11	13.1050	11.0000	9.5353	8.3810	7.5166	3.9033	3.9139	3.9411	3.9938	4.0362
12	13.7285	11.5549	9.9912	8.7988	7.7886	4.0793	4.0717	4.0998	4.1497	4.2059
13	14.2379	11.9539	10.4070	9.1365	8.1557	4.2483	4.2677	4.2983	4.3479	4.3922
14	14.7922	12.4266	10.8329	9.5350	8.4873	4.4186	4.4336	4.4744	4.5344	4.5818
15	15.2904	12.8073	11.1753	9.8687	8.7749	4.5999	4.5968	4.6167	4.6743	4.7455

Table 2(a). Critical Values for Nested Model Comparisons Using ENCNEW in Rolling Regressions. Panel B. 5% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	5.2106	4.4037	3.8175	3.3653	2.9672	1.7212	1.7250	1.7277	1.7578	1.7867
2	7.1941	6.0870	5.2817	4.6900	4.1940	2.4460	2.4524	2.4667	2.4809	2.5147
3	8.6766	7.3757	6.4414	5.6956	5.1017	2.9878	2.9942	3.0145	3.0291	3.0638
4	9.9801	8.4269	7.3542	6.5057	5.8203	3.3736	3.3821	3.4083	3.4380	3.4902
5	11.0899	9.3779	8.2062	7.2369	6.4568	3.7636	3.7636	3.7964	3.8315	3.8851
6	12.0293	10.2038	8.9327	7.8604	7.0478	4.0740	4.0749	4.1175	4.1443	4.2154
7	12.9684	10.9771	9.6020	8.4979	7.6047	4.3889	4.4147	4.4481	4.4866	4.5643
8	13.8311	11.7098	10.1977	9.0788	8.1170	4.6712	4.6758	4.7101	4.7572	4.8184
9	14.5854	12.3429	10.8114	9.5442	8.5896	4.9465	4.9664	4.9899	5.0261	5.1008
10	15.4082	13.0137	11.3283	10.0612	9.0527	5.1798	5.2039	5.2468	5.2887	5.3779
11	16.0986	13.6306	11.8735	10.4781	9.4248	5.3868	5.3977	5.4650	5.5109	5.5710
12	16.7878	14.1765	12.3970	10.9857	9.8252	5.6145	5.6394	5.6592	5.7062	5.7830
13	17.4795	14.6829	12.8751	11.3870	10.2301	5.8174	5.8200	5.8896	5.9445	6.0393
14	18.0793	15.2880	13.3657	11.8226	10.6058	6.0631	6.0683	6.1094	6.1913	6.2594
15	18.7774	15.8456	13.7500	12.1875	10.9393	6.2622	6.2559	6.3083	6.3868	6.4720

Table 2(a). Critical Values for Nested Model Comparisons Using ENCNEW in Rolling Regressions. Panel C. 1% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	8.1248	7.0317	6.2526	5.6305	5.0886	3.4345	3.4475	3.4516	3.4753	3.5225
2	10.7102	9.1525	8.1140	7.3778	6.7193	4.3709	4.4253	4.4959	4.5238	4.5651
3	12.6148	10.7701	9.5087	8.6335	7.8188	5.0151	5.0557	5.1076	5.1608	5.2130
4	14.4512	12.3304	10.9446	9.7063	8.7498	5.5978	5.6193	5.6792	5.7664	5.8489
5	15.7483	13.5035	11.8009	10.5938	9.4934	6.0746	6.1204	6.1572	6.2721	6.3551
6	17.1316	14.6207	12.9223	11.4535	10.3722	6.6328	6.6519	6.6843	6.7477	6.8566
7	18.4058	15.6036	13.6712	12.1276	11.0177	7.0298	7.0197	7.0609	7.1523	7.2397
8	19.4893	16.5436	14.4587	12.9683	11.7467	7.4327	7.5014	7.5554	7.6790	7.7505
9	20.5251	17.5530	15.2885	13.6780	12.3658	7.8075	7.8581	7.9447	8.0477	8.1781
10	21.4156	18.2554	16.1388	14.3467	12.9096	8.1720	8.2469	8.2792	8.4031	8.4557
11	22.3654	19.0642	16.8119	14.9437	13.5102	8.5524	8.5682	8.6651	8.7453	8.8469
12	23.4042	19.9109	17.4420	15.4734	13.9833	8.8930	8.9258	8.9640	9.0394	9.1783
13	24.2346	20.6730	18.1467	16.1575	14.4604	9.1303	9.1601	9.1785	9.2868	9.4391
14	25.1145	21.4808	18.7456	16.7833	14.9291	9.4630	9.4666	9.5161	9.6433	9.7375
15	25.8071	21.9110	19.2806	17.1883	15.5342	9.8260	9.8114	9.8535	9.9536	10.0418

Notes to Table 2(a). The critical values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normal random variates.

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW in Recursive Regressions. Panel A. 10% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	2.0428	1.8830	1.7435	1.6219	1.5091	0.8622	0.8775	0.8889	0.9035	0.9145
2	3.1227	2.8651	2.6639	2.4803	2.3071	1.3159	1.3341	1.3463	1.3624	1.3886
3	3.8543	3.5597	3.3135	3.0828	2.8768	1.6625	1.6933	1.7117	1.7366	1.7641
4	4.5082	4.1321	3.8364	3.5797	3.3376	1.9164	1.9396	1.9763	2.0036	2.0513
5	5.0500	4.6473	4.3051	4.0016	3.7208	2.1657	2.1996	2.2321	2.2660	2.3092
6	5.5757	5.1252	4.7305	4.4050	4.1127	2.3704	2.4048	2.4319	2.4803	2.5176
7	6.0374	5.5414	5.1276	4.7811	4.4461	2.5437	2.5800	2.6205	2.6861	2.7341
8	6.4764	5.9482	5.5127	5.1154	4.7680	2.7313	2.7723	2.8313	2.8637	2.9257
9	6.8944	6.3558	5.8931	5.4766	5.1087	2.9325	2.9670	3.0143	3.0573	3.1207
10	7.2925	6.7043	6.1977	5.7599	5.3639	3.0651	3.1145	3.1628	3.2137	3.2765
11	7.6143	7.0246	6.5176	6.0541	5.6192	3.2109	3.2607	3.3109	3.3586	3.4171
12	7.9420	7.2955	6.7881	6.3120	5.8538	3.3502	3.3884	3.4376	3.4948	3.5555
13	8.2883	7.6221	7.0714	6.5759	6.1197	3.5006	3.5339	3.5941	3.6482	3.7235
14	8.6058	7.9305	7.3606	6.8337	6.3895	3.6514	3.7110	3.7557	3.8150	3.8825
15	8.8857	8.1370	7.5860	7.0482	6.5921	3.7525	3.8152	3.8827	3.9428	4.0206

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW in Recursive Regressions. Panel B. 5% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
	$\underline{\mu}$					$\underline{\mu}$				
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	3.0638	2.8140	2.6240	2.4461	2.2837	1.4557	1.4647	1.4803	1.4944	1.5156
2	4.3131	3.9726	3.7085	3.4525	3.2181	2.0191	2.0367	2.0747	2.0998	2.1306
3	5.2001	4.7817	4.4370	4.1523	3.8574	2.4272	2.4667	2.4985	2.5294	2.5799
4	5.9751	5.4797	5.0997	4.7538	4.4246	2.7898	2.8346	2.8649	2.9151	2.9672
5	6.6020	6.0993	5.6497	5.2538	4.8908	3.0721	3.1101	3.1594	3.2014	3.2573
6	7.2016	6.6383	6.1759	5.7537	5.3856	3.3314	3.3874	3.4495	3.5145	3.5688
7	7.7958	7.1705	6.6788	6.2057	5.8055	3.6212	3.6614	3.7128	3.7675	3.8320
8	8.3056	7.6811	7.1379	6.6678	6.2018	3.8526	3.9251	3.9847	4.0383	4.0988
9	8.8298	8.0886	7.5283	7.0132	6.5634	4.1026	4.1468	4.1963	4.2526	4.3328
10	9.2405	8.5008	7.8841	7.3616	6.8901	4.2921	4.3589	4.4014	4.4695	4.5361
11	9.5814	8.8543	8.2431	7.6692	7.2053	4.4731	4.5319	4.6075	4.6818	4.7440
12	10.0759	9.2244	8.5686	7.9979	7.4890	4.6464	4.7230	4.7902	4.8681	4.9279
13	10.4586	9.6354	8.9276	8.3086	7.7620	4.8333	4.8881	4.9780	5.0596	5.1334
14	10.8035	9.9911	9.2899	8.6125	8.0456	5.0157	5.0715	5.1384	5.2180	5.3065
15	11.1341	10.3049	9.5879	8.8894	8.2925	5.1351	5.2050	5.2704	5.3477	5.4548

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW in Recursive Regressions. Panel C. 1% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
k	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	5.6201	5.1151	4.7583	4.4755	4.1745	2.8616	2.8970	2.9269	2.9625	3.0234
2	7.2437	6.5985	6.1236	5.7072	5.3488	3.6444	3.6771	3.7370	3.7917	3.8304
3	8.4061	7.6970	7.1352	6.6393	6.2252	4.1941	4.2496	4.2845	4.3363	4.3943
4	9.5015	8.7269	8.0833	7.5067	7.0207	4.7015	4.7737	4.8230	4.8818	4.9463
5	10.2276	9.3676	8.6622	8.1127	7.5346	5.1724	5.2169	5.2603	5.3311	5.4173
6	11.0099	10.0029	9.3611	8.6827	8.1067	5.4380	5.4751	5.5475	5.6192	5.7446
7	11.7372	10.7116	9.9961	9.1691	8.6190	5.7559	5.8201	5.8908	5.9751	6.0984
8	12.3869	11.4660	10.5721	9.7422	9.1030	6.1524	6.2224	6.2895	6.3647	6.4412
9	12.9844	12.0180	11.1165	10.2776	9.6076	6.4368	6.5099	6.6060	6.7043	6.7516
10	13.5982	12.6136	11.6897	10.8368	10.0149	6.7008	6.7982	6.9033	6.9543	7.0391
11	14.1987	12.9637	12.0527	11.2496	10.5364	7.0026	7.0685	7.1484	7.2302	7.3586
12	14.6368	13.3992	12.4392	11.5945	10.8842	7.2767	7.3195	7.3934	7.5024	7.5876
13	15.0736	13.8831	12.7743	11.8981	11.2180	7.4715	7.5770	7.6601	7.7316	7.8259
14	15.6463	14.3440	13.3178	12.3353	11.6119	7.6955	7.7587	7.8581	7.9879	8.1301
15	16.1904	14.8480	13.7635	12.7435	11.9302	7.9630	8.0681	8.1428	8.2522	8.3688

Notes to Table 2(b). The critical values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normal random variates. k denotes the number of additional regressors in the nesting forecasting model.

Table 3(a). Critical Values for Regression-Based Forecast Tests in Rolling Regressions. Panel A. 10% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
	$\underline{\mu}$					$\underline{\mu}$				
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	6.0232	5.6957	5.3519	5.0603	4.7289	2.3352	2.3866	2.4360	2.4863	2.5356
2	8.7104	8.3182	7.9441	7.5702	7.1775	4.0327	4.0971	4.1615	4.2310	4.3055
3	10.9175	10.5160	10.1210	9.6689	9.2188	5.5968	5.6865	5.7731	5.8579	5.9555
4	12.8892	12.4355	11.9888	11.5234	11.0328	6.9825	7.1009	7.1930	7.2968	7.4078
5	14.7178	14.2720	13.8253	13.3149	12.8049	8.3948	8.5133	8.6385	8.7684	8.8849
6	16.5061	16.0064	15.5135	14.9935	14.4440	9.6853	9.8289	9.9636	10.1149	10.2460
7	18.2076	17.6931	17.1643	16.6285	16.0431	10.9867	11.1379	11.3069	11.4364	11.5806
8	19.8602	19.2927	18.7560	18.1639	17.5679	12.2381	12.4036	12.5735	12.7378	12.9106
9	21.4528	20.8269	20.2910	19.7234	19.0474	13.5190	13.6791	13.8281	13.9988	14.1684
10	23.0218	22.3889	21.8070	21.1754	20.5231	14.7707	14.9352	15.0980	15.3045	15.4487
11	24.4742	23.8423	23.2744	22.6039	21.9323	15.9558	16.1214	16.2999	16.4800	16.6643
12	26.0187	25.3868	24.7328	24.0608	23.3456	17.1518	17.3281	17.5026	17.6942	17.8671
13	27.5256	26.8514	26.1987	25.5138	24.8065	18.4130	18.5986	18.7856	18.9767	19.1441
14	29.0283	28.3664	27.6849	27.0006	26.2935	19.6634	19.8409	20.0041	20.2057	20.3971
15	30.4502	29.7092	29.0774	28.4029	27.6317	20.8167	21.0132	21.2140	21.4499	21.6352

Table 3(a). Critical Values for Regression-Based Forecast Test in Rolling Regressions. Panel B. 5% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	7.5564	7.2275	6.8546	6.4796	6.0972	3.1920	3.2724	3.3522	3.4358	3.5206
2	10.3769	10.0324	9.6538	9.2701	8.8118	5.1033	5.2112	5.3354	5.4318	5.5503
3	12.7735	12.3460	11.9425	11.4939	11.0302	6.7967	6.9427	7.0852	7.2406	7.3766
4	14.8321	14.3901	13.9540	13.4943	12.9867	8.2969	8.4687	8.6318	8.7958	8.9664
5	16.7614	16.2736	15.8067	15.3362	14.8067	9.7989	9.9693	10.1493	10.3354	10.5375
6	18.5950	18.0767	17.6066	17.1015	16.5707	11.2498	11.4217	11.6045	11.8109	11.9937
7	20.3976	19.8283	19.3049	18.7764	18.2241	12.6046	12.7993	12.9861	13.1972	13.3996
8	22.1168	21.5762	21.1031	20.4959	19.8974	13.9372	14.1526	14.3703	14.5839	14.8045
9	23.8522	23.2566	22.6911	22.1283	21.4786	15.2917	15.5188	15.7380	15.9483	16.1988
10	25.4795	24.9024	24.2752	23.6293	23.0278	16.5691	16.7627	17.0114	17.2422	17.4755
11	27.0374	26.4198	25.7992	25.1407	24.5003	17.8549	18.0967	18.3394	18.5560	18.8105
12	28.6076	27.9345	27.3250	26.7175	26.0002	19.1315	19.3671	19.6241	19.8921	20.1026
13	30.1707	29.5526	28.9241	28.2376	27.4890	20.4014	20.6697	20.9340	21.2116	21.4713
14	31.7441	31.0769	30.4594	29.8000	29.0350	21.7238	21.9949	22.2826	22.5831	22.8479
15	33.3115	32.6722	31.9929	31.2398	30.4675	22.9763	23.2559	23.5206	23.7834	24.0850

Table 3(a). Critical Values for Regression-Based Forecast Test in Rolling Regressions. Panel C. 1% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	10.9092	10.5765	10.2265	9.9023	9.4320	5.2912	5.4610	5.6518	5.8473	6.0006
2	14.1825	13.7016	13.3209	12.8796	12.3924	7.6681	7.8800	8.1139	8.3134	8.5258
3	16.5599	16.2327	15.8236	15.4133	14.9400	9.5437	9.7501	10.0145	10.2618	10.4837
4	19.1441	18.5812	18.1744	17.7106	17.2011	11.2408	11.4497	11.7294	12.0478	12.2695
5	21.2724	20.6837	20.2955	19.7587	19.2763	12.8595	13.1439	13.4204	13.7497	14.0333
6	23.2017	22.7202	22.2382	21.7131	21.1407	14.4503	14.7644	15.0760	15.3602	15.7306
7	25.0052	24.5921	24.0584	23.5994	23.0652	15.9616	16.3350	16.7109	16.9853	17.3829
8	26.9436	26.3662	25.9507	25.3262	24.7723	17.5904	17.8867	18.2658	18.5772	18.8987
9	28.7768	28.2601	27.7835	27.1973	26.4835	19.0634	19.4360	19.8217	20.2149	20.5634
10	30.4888	29.9409	29.3264	28.7970	28.1661	20.3842	20.7962	21.1757	21.5661	21.9883
11	32.4046	31.7953	31.2932	30.6334	30.0165	21.8198	22.2137	22.7070	23.0265	23.4243
12	34.1354	33.5681	33.0216	32.4559	31.6939	23.3545	23.7738	24.1897	24.6128	24.9903
13	35.6631	35.0345	34.5340	33.9171	33.2544	24.7177	25.1323	25.5434	25.9649	26.3638
14	37.1971	36.7807	36.3541	35.5661	34.8225	26.0787	26.4930	26.9710	27.4368	27.8999
15	38.9308	38.3865	37.8899	37.3130	36.5785	27.4027	27.8453	28.3305	28.8926	29.2522

Notes to Table 3(a). The critical values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normal random variates. p denotes the number of restrictions being tested.

Table 3(b). Critical Values for Regression-Based Forecast Tests in Recursive Regressions. Panel A. 10% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	0.8667	0.8162	0.7657	0.7152	0.6647	0.5112	0.5115	0.5117	0.5119	0.5121
2	1.7237	1.6230	1.5222	1.4215	1.3207	1.0158	1.0161	1.0165	1.0168	1.0171
3	2.5790	2.4281	2.2772	2.1263	1.9753	1.5193	1.5197	1.5201	1.5205	1.5208
4	3.4335	3.2325	3.0315	2.8305	2.6293	2.0225	2.0229	2.0233	2.0238	2.0242
5	4.2875	4.0363	3.7853	3.5340	3.2827	2.5251	2.5256	2.5261	2.5265	2.5271
6	5.1412	4.8401	4.5387	4.2374	3.9361	3.0274	3.0280	3.0285	3.0291	3.0296
7	5.9944	5.6430	5.2917	4.9403	4.5888	3.5297	3.5304	3.5310	3.5316	3.5322
8	6.8471	6.4458	6.0444	5.6429	5.2417	4.0318	4.0325	4.0332	4.0337	4.0344
9	7.7001	7.2487	6.7972	6.3457	5.8942	4.5338	4.5345	4.5353	4.5360	4.5367
10	8.5525	8.0512	7.5496	7.0479	6.5466	5.0356	5.0363	5.0372	5.0379	5.0385
11	9.4053	8.8537	8.3023	7.7503	7.1988	5.5373	5.5382	5.5389	5.5397	5.5403
12	10.2577	9.6559	9.0543	8.4526	7.8508	6.0390	6.0399	6.0407	6.0415	6.0424
13	11.1099	10.4582	9.8066	9.1547	8.5029	6.5404	6.5413	6.5422	6.5430	6.5438
14	11.9621	11.2604	10.5587	9.8566	9.1545	7.0417	7.0426	7.0436	7.0446	7.0455
15	12.8147	12.0625	11.3107	10.5588	9.8068	7.5434	7.5443	7.5453	7.5462	7.5471

Table 3(b). Critical Values for Regression-Based Forecast Test in Recursive Regressions. Panel B. 5% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
	$\underline{\mu}$					$\underline{\mu}$				
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	0.8716	0.8207	0.7701	0.7194	0.6688	0.5144	0.5147	0.5150	0.5153	0.5156
2	1.7302	1.6293	1.5285	1.4274	1.3264	1.0202	1.0207	1.0211	1.0216	1.0220
3	2.5873	2.4360	2.2850	2.1337	1.9825	1.5249	1.5254	1.5259	1.5264	1.5270
4	3.4430	3.2416	3.0405	2.8390	2.6375	2.0287	2.0293	2.0299	2.0305	2.0310
5	4.2983	4.0467	3.7954	3.5437	3.2921	2.5321	2.5328	2.5334	2.5341	2.5347
6	5.1529	4.8511	4.5498	4.2478	3.9460	3.0351	3.0359	3.0366	3.0373	3.0379
7	6.0072	5.6553	5.3037	4.9517	4.5998	3.5383	3.5391	3.5399	3.5407	3.5416
8	6.8611	6.4590	6.0574	5.6555	5.2533	4.0410	4.0417	4.0426	4.0434	4.0443
9	7.7146	7.2628	6.8108	6.3590	5.9067	4.5434	4.5444	4.5452	4.5461	4.5468
10	8.5677	8.0659	7.5638	7.0618	6.5597	5.0459	5.0469	5.0477	5.0486	5.0494
11	9.4210	8.8693	8.3169	7.7645	7.2121	5.5482	5.5490	5.5499	5.5509	5.5519
12	10.2742	9.6720	9.0699	8.4674	7.8650	6.0499	6.0510	6.0521	6.0530	6.0541
13	11.1271	10.4747	9.8223	9.1699	8.5176	6.5518	6.5531	6.5541	6.5552	6.5562
14	11.9801	11.2773	10.5753	9.8732	9.1699	7.0541	7.0553	7.0565	7.0575	7.0587
15	12.8324	12.0802	11.3277	10.5752	9.8228	7.5556	7.5569	7.5580	7.5592	7.5604

Table 3(b). Critical Values for Regression-Based Forecast Test in Recursive Regressions. Panel C. 1% Nominal Significance Level

	$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
			$\underline{\mu}$					$\underline{\mu}$		
p	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1	0.8804	0.8298	0.7788	0.7276	0.6766	0.5206	0.5209	0.5213	0.5218	0.5223
2	1.7430	1.6415	1.5402	1.4390	1.3374	1.0288	1.0294	1.0300	1.0306	1.0311
3	2.6028	2.4511	2.2994	2.1479	1.9961	1.5351	1.5359	1.5368	1.5377	1.5384
4	3.4606	3.2583	3.0570	2.8551	2.6526	2.0405	2.0414	2.0422	2.0431	2.0440
5	4.3184	4.0663	3.8143	3.5617	3.3094	2.5458	2.5468	2.5479	2.5490	2.5499
6	5.1746	4.8723	4.5697	4.2672	3.9651	3.0501	3.0512	3.0521	3.0531	3.0539
7	6.0313	5.6785	5.3257	4.9732	4.6204	3.5540	3.5549	3.5562	3.5576	3.5586
8	6.8873	6.4845	6.0811	5.6777	5.2754	4.0578	4.0592	4.0605	4.0619	4.0630
9	7.7422	7.2896	6.8361	6.3834	5.9304	4.5616	4.5629	4.5643	4.5657	4.5669
10	8.5973	8.0941	7.5902	7.0879	6.5839	5.0650	5.0665	5.0681	5.0696	5.0709
11	9.4511	8.8982	8.3444	7.7924	7.2382	5.5679	5.5695	5.5712	5.5726	5.5741
12	10.3058	9.7024	9.0983	8.4956	7.8916	6.0708	6.0723	6.0738	6.0755	6.0767
13	11.1600	10.5058	9.8532	9.1998	8.5453	6.5743	6.5755	6.5771	6.5787	6.5801
14	12.0146	11.3106	10.6077	9.9037	9.1995	7.0770	7.0785	7.0801	7.0815	7.0829
15	12.8675	12.1139	11.3602	10.6071	9.8531	7.5800	7.5817	7.5830	7.5840	7.5853

Notes to Table 3(b). The critical values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications in which Brownian motions are approximated by normalized partial sums of 10,000 standard normal random variates. p denotes the number of restrictions being tested.

Table 4. Size of Nested Model Comparison Tests — DGP1

Panel A. $\underline{\mu} = 0.05, 0.15, 0.25$**,** $h = 1$ **and** $p = 0$

	$\mathcal{R}_T^{\mathcal{E}}$ test									$\mathcal{A}_T^{\mathcal{E}}$ test								
	$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$		
T $\underline{\mu}$:	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25
50	.161	.138	.128	.093	.080	.074	.039	.025	.021	.123	.121	.120	.067	.067	.064	.017	.018	.017
100	.175	.122	.112	.098	.067	.061	.030	.012	.014	.110	.105	.104	.058	.058	.057	.011	.011	.011
200	.133	.114	.106	.070	.063	.056	.018	.014	.013	.113	.108	.104	.054	.054	.051	.010	.011	.010
500	.112	.108	.107	.058	.051	.053	.014	.011	.010	.112	.110	.108	.052	.052	.053	.010	.008	.009

Panel B. $h = 1, 4, 8$**,** $\underline{\mu} = 0.15$ **and** $p = 0$

	$\mathcal{R}_T^{\mathcal{E}}$ test									$\mathcal{A}_T^{\mathcal{E}}$ test								
	$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$		
T h:	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8
50	.136	.140	.124	.079	.085	.083	.025	.032	.035	.119	.112	.082	.063	.056	.038	.016	.012	.008
100	.113	.126	.129	.065	.070	.078	.017	.020	.030	.113	.112	.106	.060	.054	.051	.012	.010	.011
200	.109	.129	.119	.053	.070	.065	.013	.018	.020	.107	.120	.111	.056	.059	.051	.011	.011	.010
500	.108	.128	.119	.055	.066	.062	.012	.015	.014	.106	.121	.120	.054	.055	.059	.011	.009	.011

Panel C. $p = 0, 3, 5$**,** $\underline{\mu} = 0.15$ **and** $h = 1$

	$\mathcal{R}_T^{\mathcal{E}}$ test									$\mathcal{A}_T^{\mathcal{E}}$ test								
	$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$		
T p:	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5
50	.132	.121	.084	.081	.065	.036	.024	.016	.006	.121	.115	.096	.067	.057	.046	.018	.012	.009
100	.128	.126	.115	.067	.069	.056	.015	.014	.011	.114	.115	.104	.061	.056	.053	.011	.013	.014
200	.106	.112	.111	.058	.061	.056	.012	.013	.010	.106	.102	.095	.053	.053	.054	.012	.013	.012
500	.104	.102	.104	.050	.055	.052	.011	.010	.012	.097	.094	.098	.049	.047	.048	.009	.011	.011

Notes to Table 4. h is the forecast horizon, $(p+1)$ is the number of regressors in the nesting forecasting model, α is the nominal significance level. The number of Monte Carlo replications is 5,000.

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

Panel A. $\underline{\mu} = .05, .15, .25$**,** $h = 1$ **and** $p = 1$

	$\mathcal{R}_T$ test									$\mathcal{A}_T$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T $\underline{\mu}$:	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25
50	.033	.049	.060	.010	.017	.019	.000	.001	.001	.066	.078	.083	.021	.029	.031	.001	.002	.002
100	.047	.059	.066	.018	.024	.023	.002	.001	.002	.078	.083	.086	.036	.039	.040	.003	.004	.004
200	.057	.069	.073	.023	.029	.031	.002	.003	.004	.084	.088	.089	.040	.041	.040	.005	.006	.007
500	.071	.079	.082	.031	.036	.036	.004	.005	.006	.091	.093	.092	.043	.042	.044	.007	.007	.007

Panel B. $h = 1, 4, 8$**,** $\underline{\mu} = .15$ **and** $p = 1$

	$\mathcal{R}_T$ test									$\mathcal{A}_T$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T h:	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8
50	.049	.000	.001	.016	.000	.000	.000	.000	.000	.077	.000	.000	.030	.000	.000	.001	.000	.000
100	.070	.003	.000	.024	.000	.000	.004	.000	.000	.093	.017	.000	.043	.003	.000	.005	.000	.000
200	.074	.018	.003	.028	.004	.000	.003	.000	.000	.090	.038	.006	.037	.013	.001	.004	.001	.000
500	.077	.072	.023	.037	.024	.005	.007	.002	.000	.092	.072	.014	.043	.033	.004	.007	.003	.000

Panel C. $p = 1, 3, 5$**,** $\underline{\mu} = .15$ **and** $h = 1$

	$\mathcal{R}_T$ test									$\mathcal{A}_T$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T p:	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5	1	3	5
50	.044	.170	.614	.012	.071	.375	.001	.004	.051	.079	.112	.234	.030	.038	.100	.002	.004	.008
100	.056	.136	.446	.021	.058	.278	.001	.006	.070	.083	.100	.168	.034	.046	.084	.003	.004	.010
200	.071	.100	.225	.026	.049	.127	.003	.008	.030	.091	.094	.120	.041	.045	.060	.005	.005	.009
500	.075	.089	.122	.035	.040	.064	.005	.006	.012	.093	.097	.108	.041	.046	.055	.008	.008	.009

Notes to Table 5. h is the forecast horizon, p is the number of regressors in model 2, α is the nominal significance level. The number of Monte Carlo replications is 5,000.

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

Panel A. $\underline{\mu} = .05, .15, .25$**,** $h = 1$ **and** $p = 1$

	$\mathcal{R}_T^{\mathcal{W}}$ test									$\mathcal{A}_T^{\mathcal{W}}$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T $\underline{\mu}$:	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25	.05	.15	.25
50	.031	.062	.089	.005	.016	.029	.000	.000	.000	.085	.106	.120	.027	.042	.051	.001	.002	.002
100	.061	.094	.113	.020	.037	.044	.001	.001	.003	.097	.112	.122	.037	.049	.054	.001	.002	.003
200	.088	.107	.119	.036	.049	.052	.003	.005	.006	.094	.113	.120	.039	.049	.055	.003	.005	.006
500	.099	.120	.138	.043	.057	.060	.007	.008	.007	.098	.116	.131	.041	.049	.058	.004	.006	.007

Panel B. $h = 1, 4, 8$**,** $\underline{\mu} = .15$ **and** $p = 1$

	$\mathcal{R}_T^{\mathcal{W}}$ test									$\mathcal{A}_T^{\mathcal{W}}$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T h:	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8	1	4	8
50	.032	.000	.004	.006	.000	.003	.000	.000	.003	.068	.016	.001	.025	.000	.000	.001	.000	.000
100	.064	.027	.006	.023	.006	.003	.002	.000	.002	.088	.096	.040	.039	.037	.005	.005	.001	.000
200	.085	.092	.047	.037	.037	.014	.006	.005	.002	.098	.141	.111	.050	.073	.050	.008	.013	.003
500	.093	.174	.136	.043	.087	.063	.009	.020	.011	.098	.165	.163	.053	.096	.087	.011	.029	.024

Panel C. $p = 1, 3, 6$**,** $\underline{\mu} = .15$ **and** $h = 1$

	$\mathcal{R}_T^{\mathcal{W}}$ test									$\mathcal{A}_T^{\mathcal{W}}$ test								
	$\alpha = .10$			$\alpha = .05$			$\alpha = .01$			$\alpha = .10$			$\alpha = .05$			$\alpha = .01$		
T p:	1	3	6	1	3	6	1	3	6	1	3	6	1	3	6	1	3	6
50	.076	.011	.007	.026	.002	.001	.001	.000	.000	.105	.023	.011	.046	.004	.001	.008	.000	.000
100	.110	.036	.024	.047	.011	.007	.006	.001	.000	.129	.026	.018	.063	.008	.003	.011	.000	.000
200	.116	.062	.057	.059	.020	.021	.010	.002	.002	.127	.027	.019	.066	.009	.004	.013	.000	.000
500	.137	.087	.091	.072	.042	.043	.018	.009	.006	.133	.032	.020	.074	.011	.006	.016	.001	.001

Notes to Table 6. h is the forecast horizon, p is the number of restrictions being tested. α is the nominal significance level. The number of Monte Carlo replications is 5,000.

Table 7. Size properties – Fixed Window

	GW_T Test		CW_T Test	
P	$\underline{R}$=20	$\underline{R}$=30	$\underline{R}$=20	$\underline{R}$=30
Panel A. 10% Nominal Size				
100	0.1282	0.1736	0.1202	0.1264
200	0.1022	0.1382	0.1044	0.1028
500	0.0654	0.1092	0.0938	0.0924
Panel B. 5% Nominal Size				
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
Panel C. 1% Nominal Size				
100	0.0322	0.0482	0.0074	0.0122
200	0.0282	0.0366	0.0068	0.0062
500	0.0096	0.0250	0.0052	0.0078

Notes to Table 7. The table reports empirical rejection frequencies of the the GW_T test, eq. (21), implemented with $\underline{R}$=20 or 30, and $\overline{R}$=$\underline{R}$+5. It also reports empirical rejection frequencies of the CW_T test, eq. (20), implemented with the same choices of window sizes. The number of Monte Carlo replications is 5,000.

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

Panel A. No Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
d_t	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	five R
0.00	0.072	0.062	0.058	0.052	0.055	0.057	0.199	0.144
0.10	0.164	0.193	0.252	0.305	0.304	0.280	0.557	0.465
0.20	0.505	0.653	0.756	0.802	0.789	0.704	0.943	0.908

	The Proposed Tests					
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
d_t	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0.00	0.070	0.063	0.056	0.054	0.054	0.051
0.10	0.191	0.260	0.284	0.318	0.319	0.320
0.20	0.634	0.770	0.806	0.830	0.831	0.830

Table 8. Rejection Frequencies of Nested Model Comparison Tests — DGP1
Panel B. One-Time Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
τ	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	five R
0	0.071	0.063	0.058	0.052	0.055	0.057	0.199	0.145
40	0.467	0.445	0.107	0.085	0.077	0.096	0.493	0.502
80	0.860	0.925	0.902	0.232	0.207	0.232	0.975	0.959
120	0.978	0.993	0.995	0.975	0.332	0.331	1.000	0.999
160	0.997	1.000	1.000	1.000	0.980	0.400	1.000	1.000

	The Proposed Tests					
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
τ	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0	0.069	0.063	0.056	0.054	0.054	0.051
40	0.508	0.275	0.131	0.171	0.075	0.064
80	0.924	0.935	0.857	0.837	0.647	0.374
120	0.994	0.998	0.996	0.994	0.985	0.955
160	1.000	1.000	1.000	1.000	1.000	1.000

Notes to Table 8. β is the parameter value with $\beta = 0$ corresponding to the null hypothesis. τ is the break date with $\tau = 0$ corresponding to the null hypothesis. The five values of R used in the last column are $R = 20, 40, 80, 120, 160$. The nominal significance level is set to 0.05. The number of Monte Carlo replications is 5,000.

Table 9. Rejection Frequencies of Nonnested Model Comparison Tests — DGP2

Panel A. No Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
d_t	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	four R
0.5	0.050	0.045	0.045	0.043	0.042	0.042	0.175	0.130
1.0	0.975	0.983	0.974	0.921	0.772	0.427	0.986	0.992
1.5	1.000	1.000	1.000	1.000	0.998	0.893	1.000	1.000

	The Proposed Tests					
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
d_t	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0.5	0.038	0.029	0.021	0.033	0.041	0.050
1.0	0.970	0.930	0.882	0.931	0.926	0.923
1.5	1.000	1.000	1.000	1.000	1.000	1.000

Table 9. Rejection Frequencies of Nonnested Model Comparison Tests — DGP2
Panel B. One-Time Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
τ	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	four R
0	0.051	0.045	0.044	0.043	0.041	0.042	0.175	0.129
40	0.111	0.073	0.043	0.039	0.041	0.041	0.187	0.161
80	0.380	0.332	0.155	0.045	0.038	0.038	0.247	0.413
120	0.695	0.686	0.518	0.117	0.048	0.042	0.562	0.753
160	0.890	0.903	0.832	0.523	0.138	0.058	0.843	0.931

	The Proposed Tests					
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
τ	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0	0.038	0.029	0.021	0.032	0.041	0.050
40	0.063	0.031	0.022	0.026	0.034	0.045
80	0.224	0.080	0.023	0.023	0.025	0.031
120	0.552	0.304	0.140	0.073	0.050	0.039
160	0.827	0.654	0.463	0.447	0.394	0.371

Notes to Table 9. β is the parameter value with $\beta = 0$ corresponding to the null hypothesis. τ is the break date with $\tau = 0$ corresponding to the null hypothesis. The five values of R used in the last column are $R = 20, 40, 80, 120, 160$. The nominal significance level is set to 0.05. The number of Monte Carlo replications is 5,000.

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

Panel A. No Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
δ_t	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	four R
0.0	0.026	0.038	0.047	0.047	0.053	0.049	0.253	0.135
0.5	0.998	0.999	0.999	0.983	0.800	0.333	1.000	1.000
1.0	1.000	1.000	1.000	0.959	0.568	0.085	1.000	1.000

	The Proposed Tests					
	$\mathcal{R}_T^W$ test			$\mathcal{A}_T^W$ test		
δ_t	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0.0	0.027	0.037	0.038	0.043	0.048	0.049
0.5	0.985	0.985	0.971	0.994	0.985	0.975
1.0	0.865	0.868	0.806	0.984	0.967	0.935

Table 10. Rejection Frequencies of Regression-Based Tests of Predictive Ability — DGP3
Panel B. One-Time Break Alternative

	Conventional Tests							
	Tests Based on Single R						Data Mining	
τ	$R = 10$	$R = 20$	$R = 40$	$R = 80$	$R = 120$	$R = 160$	all R	four R
0	0.026	0.037	0.046	0.048	0.053	0.051	0.253	0.136
40	0.035	0.040	0.034	0.037	0.042	0.035	0.198	0.112
80	0.146	0.159	0.089	0.020	0.018	0.014	0.211	0.193
120	0.431	0.494	0.352	0.073	0.006	0.006	0.513	0.516
160	0.842	0.903	0.849	0.495	0.089	0.003	0.932	0.925

	The Proposed Tests					
	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
τ	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$	$\underline{\mu} = 0.05$	$\underline{\mu} = 0.15$	$\underline{\mu} = 0.25$
0	0.111	0.037	0.003	0.083	0.047	0.011
40	0.078	0.022	0.002	0.067	0.038	0.008
80	0.100	0.022	0.001	0.063	0.024	0.003
120	0.332	0.108	0.004	0.180	0.059	0.005
160	0.804	0.493	0.043	0.704	0.410	0.079

Notes to Table 10. δ is the parameter value with $\delta = 0$ corresponding to the null hypothesis of no predictive ability. τ denotes the break date with $\tau = 0$ indicating that there is no structural break and the null hypothesis of no predictive ability holds. 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

Note: The table shows asymptotic rejections of nominal 5% tests for nonnested models (DMW_T), forecast optimality ($\mathcal{W}_T^{(r)}$) and nested models ($ENCNEW_T$) repeated over sequences of windows sizes equal to $\left[\underline{\mu}T\right]$, $\left[\underline{\mu}T+1\right]$, ... , $\left[\left(1-\underline{\mu}\right)T\right]$. 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

Note. Two asterisks denote significance at the 5% level, and one asterisk denotes significance at the 10% level. For the $\mathcal{R}_T$ and $\mathcal{A}_T$ 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 Panel A

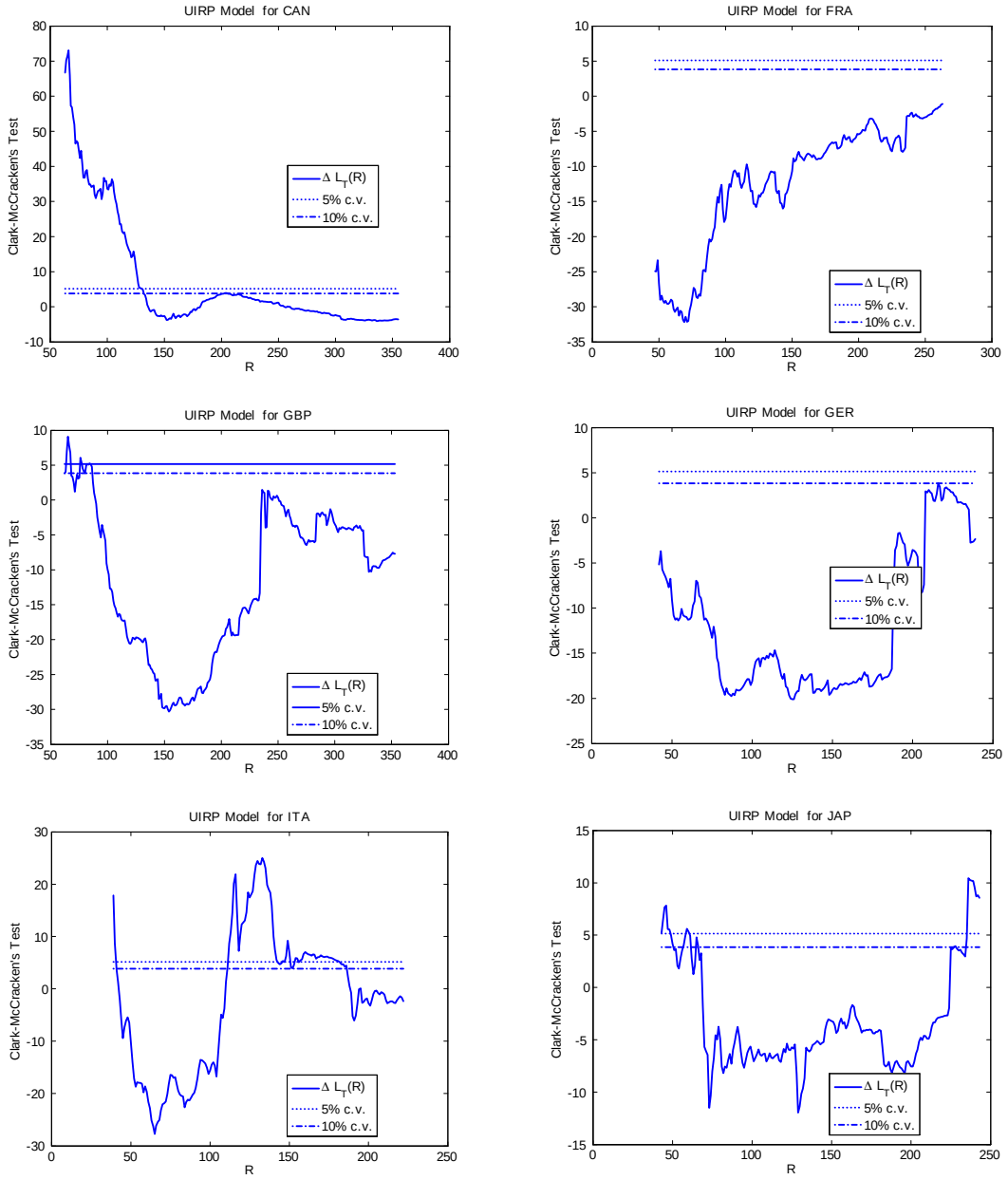

Figure 1 Panel B

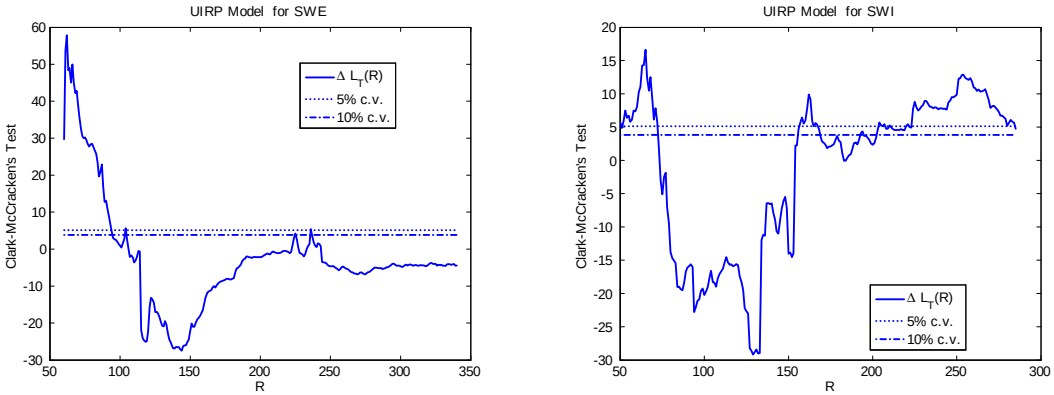

Figure 1 plots the estimated Clark and McCracken (2001) ENCNEW test statistic for comparing the UIRP model with the random walk for the window sizes we consider (reported on the x-axis), together with 5% and 10% critical values of the $\mathcal{R}_T^{\mathcal{E}}$ test statistic. The test rejects when 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), Sweden (SWE) and Switzerland (SWI).

Figure 2 Panel A

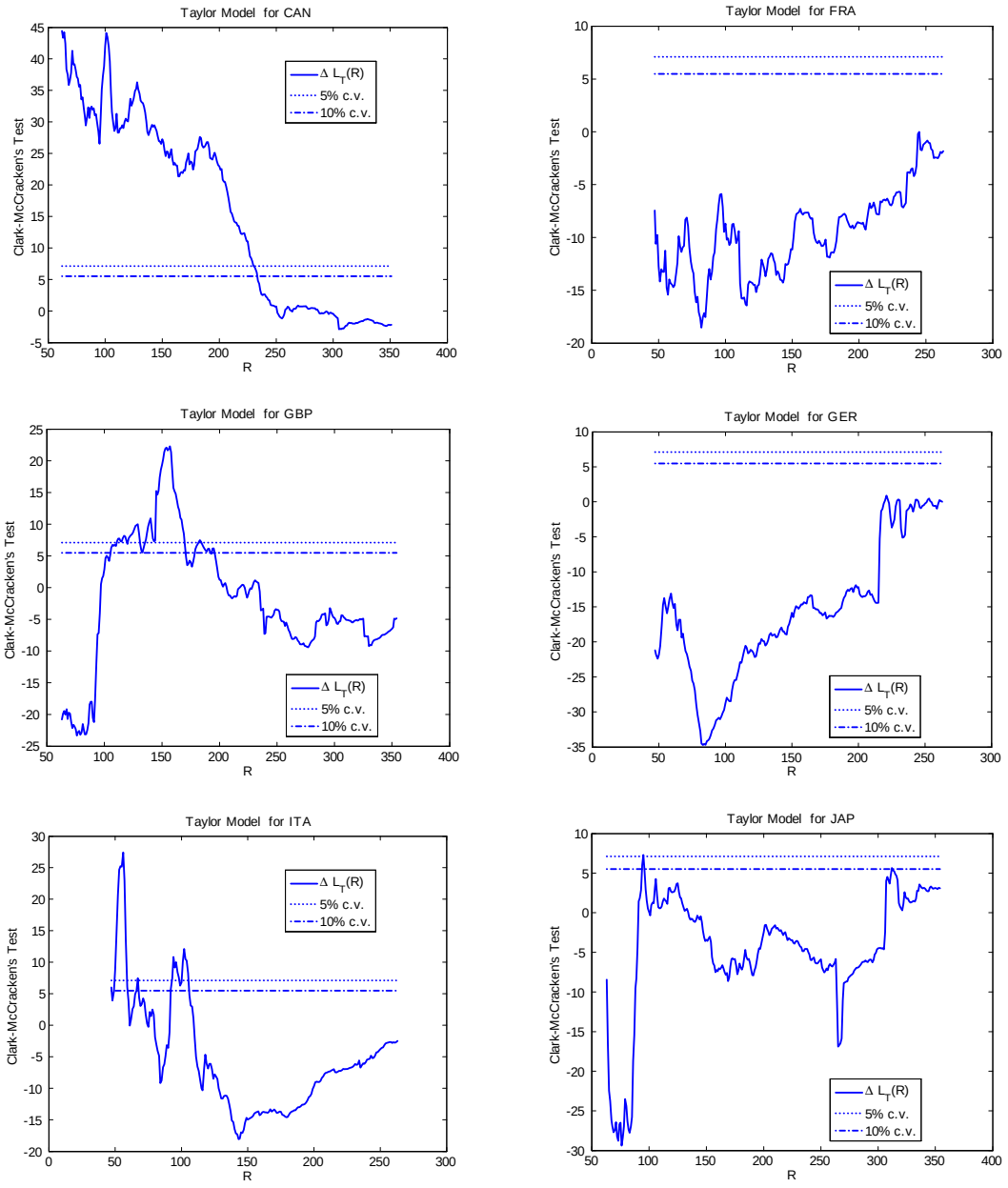

Figure 2 Panel B

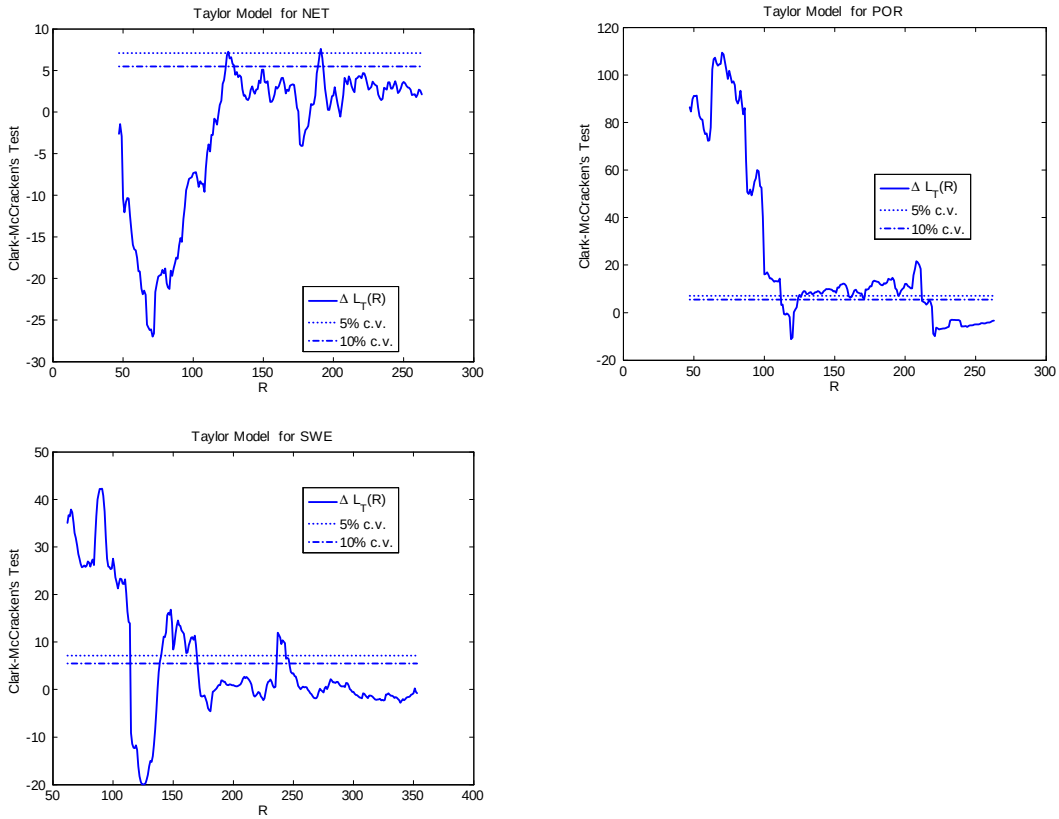

Figure 2 plots the estimated Clark and McCracken (2001) ENCNEW test statistic for comparing the Taylor rule model with the random walk for the window sizes that consider (reported on the x-axis), together with 5% and 10% critical values of the $\mathcal{R}_T^{\mathcal{E}}$ test statistic. The test rejects 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), Sweden (SWE) and Switzerland (SWI), the Netherlands (NET) and Portugal (POR).