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

Barbara Rossi and Atsushi Inoue

Duke University and NC State

January 16, 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 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 M. McCracken, J. Nason, A. Patton, K. Sill, D. Thornton and seminar participants at the Econometrics Workshop at the St. Louis Fed, Bocconi University and the Triangle Econometrics Conference for useful comments and suggestions.

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.[1] Notwithstanding the importance of the problem, no satisfactory solution has been proposed so far,[2] and in the forecasting literature it is common to only report empirical results for one window size.[3] 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.[4] 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

[1] The problem we address affects not only rolling window forecasting schemes, which we focus on, but also recursive window forecasting schemes, which we discuss in the Appendix.

[2] As discussed in 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 may not provide as precise estimation as larger windows if the data are stationary. The empirical evidence shows that instabilities are widespread; e.g. Stock and Watson (2003a) for macroeconomic data, Paye and Timmermann (2006) for asset returns and Rossi (2006) for exchange rate models.

[3] For example, see Meese and Rogoff (1983a), Chinn (1991), Qi and Wu (2003), Cheung et al. (2005), van Dijk and Frances (2005), Clark and West (2006, 2007), Gourinchas and Rey (2007), and Molodtsova and Papell (2009).

[4] 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.

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), Clark and McCracken (2001) and Clark and West (2007).[5] 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. Although the paper focuses on rolling window forecast, similar results are developed for recursive window forecasts, and are provided in the Appendix.

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 objective is to improve the model's forecast. Similarly, Clark and McCracken (2009) combine rolling and recursive *forecasts* in the attempt of improving the forecasting model. Our paper instead proposes to take a 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 know 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. 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

[5] The only assumption we make is that the window size be large relative to the sample size.

over several window sizes, and removes the arbitrary decision of the choice of the window length.[6]

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.

The paper is organized as follows. Section 2 presents the econometric methodology. Section 3 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 Econometric methodology

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 a rolling window forecast scheme. The case of the recursive window forecast scheme will be considered in Appendix B. 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. Let researchers bee interested in evaluating the forecasting performance of two competing models: Model 1, involving parameters θ, and Model 2, involving parameters γ. In the rolling window forecast method, the true but unknown model's parameters θ^* and γ^* are estimated by $\widehat{\theta}_{t,R}$ and $\widehat{\gamma}_{t,R}$, which are calculated using samples of R observations dated $t-R+1, ..., t$, for $t = R$, $R+1, ..., T$. Let $\left\{L^{(1)}_{t+h}\left(\widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$ and $\left\{L^{(2)}_{t+h}\left(\widehat{\gamma}_{t,R}\right)\right\}_{t=R}^{T}$ 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.[7]

[6] See White (2000) or Clark and McCracken (2010) for techniques robust to data snooping in a very different context, namely multiple model comparisons.

[7] The rolling scheme case involves in principle the choice of two nuisance parameters: the size of the estimation window, and the first estimation window, which could in principle be different. In practice,

We start by discussing results pertaining 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' (1969) regressions, among others.

2.1 Non-Nested Model Comparisons

Let the researcher evaluate the two models using the sample average of the sequence of standardized out-of-sample loss differences:

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

where $\widehat{\sigma}_R^2$ is a consistent estimate of the long run variance matrix of the out-of-sample loss differences. For example, this is the strategy adopted by Diebold and Mariano (1995), West (1996) and McCracken (2000).[8] The test statistics that each of these papers propose differ in the estimate of the variance. Since our approach is valid no matter which test statistic is considered, we will defer details on how the variance estimate is constructed.

We make the following high level assumption.[9]

Assumption 1:

(a) The partial sum $\hat{\sigma}_R^{-1} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \{\Delta L_{t+h}(\hat{\theta}_{t,R}, \hat{\gamma}_{t,R}) - E[\Delta L_{t+h}(\theta^*, \gamma^*)]\}$ *obeys a functional central limit theorem:*

$$\hat{\sigma}_R^{-1} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \left\{ \Delta L_{t+h}(\hat{\theta}_{t,R}, \hat{\gamma}_{t,R}) - E[\Delta L_{t+h}(\theta^*, \gamma^*)] \right\} \Rightarrow \mathcal{B}(\cdot) - \mathcal{B}(\mu), \tag{1}$$

where $[x]$ *denotes the integer part of* x, $\Rightarrow$ *denotes weak convergence on the space of cadlag functions on* $(0, 1)$, $D(0, 1)$ *equipped with the Skorohod topology, and* $B(\cdot)$ *denotes the standard Brownian motion (see Billingsley (1968) for the definitions of weak convergence, space* D *and the Skorohod topology)*;

however, researchers choose the first estimation window to be the same as the size of the rolling window, the case we focus on. Appendix B shows that our proposed tests can be extended to the recursive scheme in a straightforward manner. We focus on the rolling scheme in the main text because it is the leading case in which the value of R is manipulated in practice.

[8] Note that $\left(\frac{T}{P}\right)^{1/2} \Delta L_T\left(R\right)$ would be exactly the test statistic proposed by Diebold and Mariano (1995), West (1996) and McCracken (2000).

[9] See Rossi and Sekhposyan (2010) for more primitive assumptions.

(b) $lim_{T,R\rightarrow\infty}R/T=\mu\in(0,1)$.

Assumption 1 implies $\widehat{\sigma}_R^{-1}P^{-1/2}\sum_{t=R}^{T}\Delta L_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R})$ is asymptotically normally distributed for a given estimation window R, and it is therefore satisfied for the Diebold and Mariano (1995), West (1996) and McCracken (2000) tests, under their assumptions.

Proposition 1 describes our proposed procedure for non-nested models. 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, and corresponds to the test labeled $\mathcal{R}_T$. 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 is to take a weighted average of the $\Delta L_T(R)$ tests, giving equal weight to each test. This would correspond to the test labeled $\mathcal{A}_T$.

Proposition 1 (Out-of-sample Robust test for Non-nested Models) *Suppose Assumption 1 holds. Let*

$$\mathcal{R}_T=\sup_R|\Delta L_T(R)|,\ \ R=\underline{R},...\overline{R},\tag{2}$$

and

$$\mathcal{A}_T=\frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}}|\Delta L_T(R)|,\tag{3}$$

where

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

$R=[\mu T]$, $\underline{R}=[\underline{\mu}T]$, $\overline{R}=[\overline{\mu}T]$, *and* $\hat{\sigma}_R^2$ *is a consistent estimator of* σ^2.
Under the null hypothesis $H_0:\lim_{T\rightarrow\infty}E[\Delta L_T^*(R)]=0$ *for all* R,

$$\mathcal{R}_T\Longrightarrow\sup_{\mu\in[\underline{\mu},\overline{\mu}]}|\mathcal{B}(1)-\mathcal{B}(\mu)|,\tag{4}$$

and

$$\mathcal{A}_T\Longrightarrow\int_{\underline{\mu}}^{\overline{\mu}}|\mathcal{B}(1)-\mathcal{B}(\mu)|\,d\mu,\tag{5}$$

where $\mathcal{B}(\cdot)$ *is a standard univariate Brownian motion. The null hypothesis for the* $\mathcal{R}_T$ *test is rejected at the significance level* α *when* $\mathcal{R}_T>k_\alpha^{\mathcal{R}}$ *whereas the null hypothesis for the* $\mathcal{A}_T$ *test is rejected when* $\mathcal{A}_T>k_\alpha^{\mathcal{A}}$, *where the critical values* $\left(\alpha,k_\alpha^{\mathcal{R}}\right)$ *and* $\left(\alpha,k_\alpha^{\mathcal{A}}\right)$ *for various values of* $\underline{\mu}$ *and* $\overline{\mu}=1-\underline{\mu}$ *are reported in Table 1.*

The proof of Proposition 1 is provided in Appendix A. 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

$$P\left\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]}|\mathcal{B}(1)-\mathcal{B}(\mu)|>k_\alpha^{\mathcal{R}}\right\} = \alpha \tag{6}$$

$$P\left\{\int_{\underline{\mu}}^{\overline{\mu}}|\mathcal{B}(1)-\mathcal{B}(\mu)|\,d\mu>k_\alpha^{\mathcal{A}}\right\} = \alpha, \tag{7}$$

and are computed using Monte Carlo simulation methods.

Note also that the practical implementation of (2) and (3) requires researchers to choose $\underline{R}$ and $\overline{R}$. To avoid data snooping over the choices of $\underline{R}$ and $\overline{R}$, we recommend researchers to fix $\overline{R}=T-\underline{R}$, and to use $\underline{R}=[0.15T]$ in practice.

A consistent estimate of σ^2 for non nested model comparisons in rolling windows is provided by McCracken (2000, p. 203, eqs. 5 and 6). For example, a consistent estimator when parameter estimation error is negligible is:

$$\hat{\sigma}_R^2=\sum_{i=-q(P)+1}^{q(P)-1}(1-|i/q(P)|)P^{-1}\sum_{t=R}^{T}\Delta L_{t+h}^d\left(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R}\right)\Delta L_{t+h-i}^d\left(\widehat{\theta}_{t-i,R},\widehat{\gamma}_{t-i,R}\right), \tag{8}$$

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 (8) can be used is when the same loss function is used for estimation and evaluation.

INSERT TABLE 1 HERE

2.2 Nested Models Comparison

For the case of nested models comparison, we follow Clark and McCracken (2001) and Clark and West (2007), and focus on the empirically leading case of a quadratic loss function, which implies that the predictive ability is evaluated according to the Mean Squared Forecast Error.[10] Let Model 1 be the parsimonious model, and Model 2 be the larger model that nests Model 1. Let y_{t+h} the period t forecasts of y_{t+h} from the two models be denoted by $\widehat{y}_{1t,t+h}$ and $\widehat{y}_{2t,t+h}$. We consider two separate test statistics, proposed by Clark and McCracken (2001) and Clark and West (2007), respectively.

[10] The reason we focus on quadratic loss functions is that this is a crucial assumption in Clark and West (2007), which allows them to obtain asymptotically normal test statistics. Since the methods proposed in this paper rely on the asymptotic normality result, we maintain their assumption.

For the latter, define the adjusted mean square prediction error as:

$$\Delta L^{adj}_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R}) = (y_{t+h} - \widehat{y}_{1t,t+h})^2 - \left[(y_{t+h} - \widehat{y}_{2t,t+h})^2 - (\widehat{y}_{1t,t+h} - \widehat{y}_{2t,t+h})^2\right],$$

so that the rescaled Clark and West (2007) test statistic is:[11]

$$\Delta L^{adj}_{T}(R) \equiv \frac{1}{\widehat{\sigma}_R} T^{-1/2} \sum_{t=R}^{T} \Delta L^{adj}_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R}), \tag{9}$$

where $\widehat{\sigma}^2_R$ is a consistent estimate of the long run variance matrix of the adjusted out-of-sample loss differences, $\Delta L^{adj}_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R})$. Note that, since the models are nested, the Clark and West's (2007) is one sided; as a consequence, our test procedure will be one sided too. Clark and West (2007) show that the test statistic (9) is "approximately normal"; if we extend the Clark and West's (2007) argument of "approximate normality" to partial sums of rescaled adjusted out-of-sample loss differences, under their assumptions a logical approximation can be obtained as follows.

We make the following high level assumption.

Assumption 2:

(a) The partial sum $\hat{\sigma}_R^{-1} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \left\{ \Delta L^{adj}_{t+h}(\hat{\theta}_{t,R},\hat{\gamma}_{t,R}) - E[\Delta L^{adj}_{t+h}(\theta^*,\gamma^*)] \right\}$ *obeys a functional central limit theorem:*

$$\hat{\sigma}^{-1} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \left\{ \Delta L^{adj}_{t+h}(\hat{\theta}_{t,R},\hat{\gamma}_{t,R}) - E[\Delta L^{adj}_{t+h}(\theta^*,\gamma^*)] \right\} \Rightarrow \mathcal{B}(\cdot) - \mathcal{B}(\mu); \tag{10}$$

(b) $lim_{T,R\to\infty} R/T = \mu \in (0,1)$.

Assumption 2 implies $\widehat{\sigma}_R^{-1} P^{-1/2} \sum_{t=R}^{T} \Delta L^{adj}_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R})$ is approximately asymptotically normally distributed for a given estimation window R, as in Clark and West (2007).[12]

Proposition 2 presents our robust test for the case of nested models comparisons.

[11] The original Clark and West's (2007) statistic is $\frac{1}{\widehat{\sigma}_R} P^{-1/2} \sum_{t=R}^{T} \Delta L^{adj}_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R})$.

[12] More precisely, Clark and West (2007) numerically show that the tail of the asymptotic distribution of their test statistic can be bounded by that of the standard normal distribution. Because normal approximations to the distribution of their test statistic yields a conservative test for given R, our test is also conservative which is confirmed by the Monte Carlo experiment. Note that Clark and McCracken's (2001) ENCNEW test would not satisfy this assumption, since it does not have an asymptotically Normal distribution.

Proposition 2 (Out-of-sample Robust test for Nested Models) *Suppose Assumption 2 holds. Let*

$$\mathcal{R}_T = \sup_R \Delta L_T^{adj}(R), \;\; R = \underline{R}, ...\overline{R}, \tag{11}$$

and

$$\mathcal{A}_T = \frac{1}{\overline{R} - \underline{R} + 1} \sum_{R=\underline{R}}^{\overline{R}} \Delta L_T^{adj}(R), \tag{12}$$

where

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

and $\hat{\sigma}_R^2$ is a consistent estimator of σ^2. Under the null hypothesis $H_0 : \lim_{T\to\infty} E[\Delta L_T^{adj}(R)] = 0$ for all R,

$$\mathcal{R}_T \Longrightarrow \sup_{\mu \in [\underline{\mu}, \overline{\mu}]} [\mathcal{B}(1) - \mathcal{B}(\mu)], \tag{13}$$

and

$$\mathcal{A}_T \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} [\mathcal{B}(1) - \mathcal{B}(\mu)] \, d\mu, \tag{14}$$

where $R = [\mu T]$, $\underline{R} = [\underline{\mu} T]$, $\overline{R} = [\overline{\mu} T]$, and $\mathcal{B}(\cdot)$ is a standard univariate Brownian motion. The null hypothesis for the $\mathcal{R}_T$ test is rejected at the significance level α when $\mathcal{R}_T > k_\alpha^{\mathcal{R}}$ whereas the null hypothesis for the $\mathcal{A}_T$ test is rejected when $\mathcal{A}_T > k_\alpha^{\mathcal{A}}$, where the critical values $(\alpha, k_\alpha^{\mathcal{R}})$ and $(\alpha, k_\alpha^{\mathcal{A}})$ for various values of $\underline{\mu}$ and $\overline{\mu} = 1 - \underline{\mu}$ are reported in Table 2(a).

The proof of Proposition 2 is provided in Appendix A. 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

$$P\left\{ \sup_{\mu \in [\underline{\mu}, \overline{\mu}]} [\mathcal{B}(1) - \mathcal{B}(\mu)] > k_\alpha^{\mathcal{R}} \right\} = \alpha \tag{15}$$

$$P\left\{ \int_{\underline{\mu}}^{\overline{\mu}} [\mathcal{B}(1) - \mathcal{B}(\mu)] \, d\mu > k_\alpha^{\mathcal{A}} \right\} = \alpha. \tag{16}$$

The critical values are obtained via Monte Carlo simulation methods.

A consistent estimator of σ^2 when parameter estimation error is negligible is

$$\hat{\sigma}_R^2 = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{t=R}^{T} \Delta L_{t+h}^{adj,d}\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) \Delta L_{t+h-i}^{adj,d}\left(\widehat{\theta}_{t-i,R}, \widehat{\gamma}_{t-i,R}\right), \tag{17}$$

where $\Delta L_{t+h}^{adj,d}\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) \equiv \Delta L_{t+h}^{adj}\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right) - P^{-1} \sum_{t=R}^{T} \Delta L_{t+h}^{adj}\left(\widehat{\theta}_{t,R}, \widehat{\gamma}_{t,R}\right)$ and $q(P)$ is a bandwidth that grows with P (e.g., Newey and West, 1987)

INSERT TABLE 2a HERE

Clark and West's (2007) test is only approximately asymptotically normal, and it may be conservative. Thus, we also consider statistics based on the Clark and McCracken's (2001) ENCNEW test, which may have better small sample properties. Define the ENCNEW statistic as:

$$\Delta L_T^{\mathcal{E}}(R) \equiv P\frac{P^{-1}\sum_{t=R}^{T}\left[\left(y_{t+h}-\widehat{y}_{1t,t+h}\right)^2-\left(y_{t+h}-\widehat{y}_{1t,t+h}\right)\left(y_{t+h}-\widehat{y}_{2t,t+h}\right)\right]}{P^{-1}\sum_{t=R}^{T}\left(y_{t+h}-\widehat{y}_{2t,t+h}\right)^2}, \tag{18}$$

where P is the number of out-of-sample predictions available, and $\widehat{y}_{1t,t+h}, \widehat{y}_{2t,t+h}$ depend on the parameter estimates $\hat{\theta}_{t,R}, \hat{\gamma}_{t,R}$.

As in Clark and McCracken (2001), we make the following assumptions..[13]

Assumption 3:

(a) The parameter estimates $\hat{\theta}_{t,R}$ satisfies $\hat{\theta}_{t,R}-\theta^* = B_1(t)\Psi_1(t)$ where $B_1(t)\Psi_1(t) = \left(R^{-1}\sum_{j=t-R+1}^{t}q_{1,j}\right)^{-1}\left(R^{-1}\sum_{j=t-R+1}^{t}\psi_{1,j}\right)$, and similarly for $\hat{\gamma}_{t,R}-\gamma^* = B_2(t)\Psi_2(t)$.

(b) Let $U_t = \left[u_t, x_{2,t}'-Ex_{2,t}', \psi_{2,t}', vec\left(\psi_{2,t}\psi_{2,t}'-E\psi_{2,t}\psi_{2,t}'\right)', vec\left(q_{2,t}-Eq_{2,t}\right)'\right]'$. *Then* $EU_t = 0$; $Eq_{2,t} < \infty$ *is p.d.; for some* $r > 4$, U_t *is uniformly* L^r *bounded; for all t,* $Eu_t^2 = \sigma^2 < \infty$; *for some* $r > d > 2$, U_t *is strong mixing with coefficients of size* $-rd/(r-d)$; *letting* $\widetilde{U}_t$ *denote the vector of non-redundant elements of* U_t, $lim_{T\to\infty}T^{-1}E\left(\sum_{t=1}^{T}\widetilde{U}_t\right)\left(\sum_{t=1}^{T}\widetilde{U}_t\right)' = \Omega < \infty$ *is p.d.*

(c) $E\left(\psi_{2,t}\psi_{2,t}'\right) = \sigma^2 Eq_{2,t}$ *and* $E\left(\psi_{2,t}|\psi_{2,t-j}, q_{2,t-j,}j = 1,2,...\right) = 0$.

(d) $lim_{T,R\to\infty}R/T = \mu \in (0,1)$ *and the forecast horizon is one.*

Note that Assumption 3 imposes conditional homoskedasticity and one-step ahead forecast horizons; for situations where conditional heteroskedasticity and multi-step predictions are important, see Clark and McCracken (2005b). Proposition 3 presents our robust test for the case of nested models comparisons. The proof of Proposition 3 is provided in Appendix A.

Proposition 3 (Out-of-sample Robust test for Nested Models II) *Suppose Assumption 3 holds. Let*

$$\mathcal{R}_T^{\mathcal{E}} = \sup_R \Delta L_T^{\mathcal{E}}(R), \; R = \underline{R}, ...\overline{R}, \tag{19}$$

[13] See Clark and McCracken (2001) for a discussion of these assumptions.

and

$$\mathcal{A}_T^{\mathcal{E}} = \frac{1}{\overline{R} - \underline{R} + 1} \sum_{R=\underline{R}}^{\overline{R}} \Delta L_T^{\mathcal{E}}(R), \tag{20}$$

where $\Delta L_T^{\mathcal{E}}(R)$ *is defined in (18). Under the null hypothesis* $H_0 : \lim_{T\to\infty} E[\Delta L_T^{\mathcal{E}}(R)] = 0$ *for all* R,

$$\mathcal{R}_T^{\mathcal{E}} \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]} \mu^{-1} \int_{\mu}^{1} [\mathcal{B}_k(s) - \mathcal{B}_k(s-\mu)]' \, d\mathcal{B}_k(s), \tag{21}$$

and

$$\mathcal{A}_T^{\mathcal{E}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left\{ \mu^{-1} \int_{\mu}^{1} [\mathcal{B}_k(s) - \mathcal{B}_k(s-\mu)]' \, d\mathcal{B}_k(s) \right\}, \tag{22}$$

where $\underline{R} = [\underline{\mu}T]$, $\overline{R} = [\overline{\mu}T]$, *and* $\mathcal{B}_k(\cdot)$ *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. The null hypothesis for the* $\mathcal{R}_T^{\mathcal{E}}$ *test is rejected at the significance level* α *when* $\mathcal{R}_T^{\mathcal{E}} > k_\alpha^{\mathcal{R}}$ *whereas the null hypothesis for the* $\mathcal{A}_T^{\mathcal{E}}$ *test is rejected when* $\mathcal{A}_T^{\mathcal{E}} > k_\alpha^{\mathcal{A}}$, *where the critical values* $\left(\alpha, k_\alpha^{\mathcal{R}}\right)$ *and* $\left(\alpha, k_\alpha^{\mathcal{A}}\right)$ *for various values of* $\underline{\mu}$ *and* $\overline{\mu} = 1 - \underline{\mu}$ *are reported in Table 2(b).*[14]

INSERT TABLE 2(b) HERE

2.3 Regression-Based Tests of Predictive Ability

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

[14] 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 $P\left\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]} \mu^{-1} \int_{\mu}^{1} [\mathcal{B}_k(s) - \mathcal{B}_k(s-\mu)]' \, d\mathcal{B}_k(s) \ > k_\alpha^{\mathcal{R}}\right\} = \alpha$ and $P\left\{\int_{\underline{\mu}}^{\overline{\mu}} \left\{\mu^{-1} \int_{\mu}^{1} [\mathcal{B}_k(s) - \mathcal{B}_k(s-\mu)]' \, d\mathcal{B}_k(s)\right\} \ > k_\alpha^{\mathcal{A}}\right\} = \alpha$, respectively. The critical values are obtained via Monte Carlo simulation methods.

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

(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).

(iii) *Forecast Encompassing Tests* (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 to be the $(p \times 1)$ vector $\mathcal{L}_{t+h} = v_{t+h} g_t$, 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 $g_t(\widehat{\theta}_t) \equiv \widehat{g}_t$.[15] The null hypothesis of interest is:

$$E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0. \tag{23}$$

In order to test (23), one simply tests whether $\widehat{\mathcal{L}}_{t+h}$ has zero mean by a standard Wald test in a regression of $\widehat{\mathcal{L}}_{t+h}$ onto a constant (i.e. testing whether the constant is zero). That is,

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

where $\widehat{\Omega}_R$ is a consistent estimate of the long run variance matrix of the adjusted out-of-sample losses discussed in details below.

We make the following high level assumption.

Assumption 4: (a) The partial sum $\hat{\Omega}^{-\frac{1}{2}} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \left\{ \mathcal{L}_{t+h}(\hat{\theta}_{t,R}) - E[\mathcal{L}_{t+h}(\theta^*)] \right\}$ *obeys a functional central limit theorem:*

$$\hat{\Omega}_R^{-\frac{1}{2}} T^{-1/2} \sum_{t=[\mu T]}^{[\cdot T]} \left\{ \mathcal{L}_{t+h}(\hat{\theta}_{t,R}) - E[\mathcal{L}_{t+h}(\theta^*)] \right\} \Rightarrow \mathcal{B}_p(\cdot) - \mathcal{B}_p(\mu) \tag{24}$$

where $B_p(\cdot)$ *denotes the p-dimensional standard Brownian motion; (b)* $\lim_{T,R\to\infty} R/T = \mu \in (0,1)$.

We have the following proposition. The proof of Proposition 4 is provided in Appendix A.[16]

[15] In the examples above, we have: (i) $g_t = 1$; (ii) $g_t = X_t$; (iii) $g_t = f_t$; (iv) $g_t = v_t$.

[16] 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 tests is: $\mathcal{W}_T(R) = P^{-1} \sum_{t=R}^{T} \widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})' \widehat{\Omega}_R^{-1} \sum_{t=R}^{T} \widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})$.

Proposition 4 (Robust Regression-Based Tests) *Suppose Assumption 4 holds. Let*

$$\mathcal{R}_T^{\mathcal{W}} = \sup_R [\mathcal{L}_T(R)' \, \widehat{\Omega}_R^{-1} \mathcal{L}_T(R)], \;\; R = \underline{R}, ...\overline{R}, \tag{25}$$

and

$$\mathcal{A}_T^{\mathcal{W}} = \frac{1}{\overline{R} - \underline{R} + 1} \sum_{R=\underline{R}}^{\overline{R}} [\mathcal{L}_T(R)' \, \widehat{\Omega}_R^{-1} \mathcal{L}_T(R)], \tag{26}$$

where

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

and $\widehat{\Omega}_R$ *is a consistent estimator of* Ω*. Under the null hypothesis* $H_0 : \lim_{T\to\infty} E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0$ *for all* R*,*

$$\mathcal{R}_T^{\mathcal{W}} \Longrightarrow \sup_{\mu \in [\underline{\mu}, \overline{\mu}]} [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)]' [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)], \tag{27}$$

and

$$\mathcal{A}_T^{\mathcal{W}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)]' [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)] \, d\mu, \tag{28}$$

where $R = [\mu T]$, $\underline{R} = [\underline{\mu} T]$, $\overline{R} = [\overline{\mu} T]$, *and* $\mathcal{B}_p(\cdot)$ *is a standard p-dimensional Brownian motion. The null hypothesis for the* $\mathcal{R}_T^{\mathcal{W}}$ *test is rejected at the significance level* α *when* $\mathcal{R}_T^{\mathcal{W}} > k_\alpha^{\mathcal{R}}$ *whereas the null hypothesis for the* $\mathcal{A}_T^{\mathcal{W}}$ *test is rejected when* $\mathcal{A}_T^{\mathcal{W}} > k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$*, where the critical values* $\left(\alpha, k_{\alpha,p}^{\mathcal{R},\mathcal{W}}\right)$ *and* $\left(\alpha, k_{\alpha,p}^{\mathcal{A},\mathcal{W}}\right)$ *for various values of* $\underline{\mu}$ *and* $\overline{\mu} = 1 - \underline{\mu}$ *are reported in Table 3.*[17]

INSERT TABLE 3 HERE

A simple, consistent estimator for $\widehat{\Omega}$ that ignores parameter estimation uncertainty is:

$$\widehat{\Omega}_R = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{t=R}^{T} \widehat{\mathcal{L}}_{t+h}^{(d)}(\widehat{\theta}_{t,R}) \widehat{\mathcal{L}}_{t+h-i}^{(d)}(\widehat{\theta}_{t-i,R})',$$

where $\widehat{\mathcal{L}}_{t+h}^{(d)}(\widehat{\theta}_{t,R}) \equiv \widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R}) - P^{-1} \sum_{t=R}^{T} \widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})$ and $q(P)$ is a bandwidth that grows with P (e.g., Newey and West, 1987). West and McCracken (1998) have instead proposed a general variance estimator that takes into account estimation uncertainty. See West and McCracken (1998) for conditions under which parameter estimation uncertainty is irrelevant.

[17] Again, the 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 $P\left\{\sup_{\mu \in [\underline{\mu}, \overline{\mu}]} [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)]' [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)] > k_\alpha^{\mathcal{R},\mathcal{W}}\right\} = \alpha$, $P\left\{\int_{\underline{\mu}}^{\overline{\mu}} [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)]' [\mathcal{B}_p(1) - \mathcal{B}_p(\mu)] \, d\mu > k_\alpha^{\mathcal{A},\mathcal{W}}\right\} = \alpha$.

Historically, researchers have estimated the alternative regression: $\widehat{v}_{t+h}=\widehat{g}_t^{\prime}\cdot\widehat{\alpha}\left(R\right)+\widehat{\eta}_{t+h}$, where $\widehat{\alpha}\left(R\right)=\left(P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{g}_t^{\prime}\right)^{-1}\left(P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{v}_{t+h}\right)$ 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 $E\left(g_t g_t^{\prime}\right)$ is full rank, a maintained assumption in that literature, then 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:[18]

$$\mathcal{W}_T^{(r)}\left(R\right)=\frac{P}{T}\widehat{\alpha}\left(R\right)^{\prime}\ \widehat{V}_{\alpha}^{-1}(R)\widehat{\alpha}\left(R\right),$$

where $\widehat{V}_{\alpha}(R)$ is a consistent estimate of the asymptotic variance of $\widehat{\alpha}\left(R\right)$, V_{α}. Proposition 5 presents results for this statistic. The proof of Proposition 5 is provided in Appendix A.

Proposition 5 (Robust Regression-Based Tests II) *Suppose Assumption 4 holds and $E\left(g_t g_t^{\prime}\right)$ is full rank. Let*

$$\mathcal{R}_T^{\mathcal{W}}=\sup_R\frac{P}{T}\widehat{\alpha}\left(R\right)^{\prime}\ \widehat{V}_{\alpha}^{-1}(R)\widehat{\alpha}\left(R\right),\ R=\underline{R},...\overline{R}, \tag{29}$$

and

$$\mathcal{A}_T^{\mathcal{W}}=\frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}}\frac{P}{T}\widehat{\alpha}\left(R\right)^{\prime}\ \widehat{V}_{\alpha}^{-1}(R)\widehat{\alpha}\left(R\right), \tag{30}$$

where

$$\widehat{\alpha}\left(R\right)=\left(P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{g}_t^{\prime}\right)^{-1}\left(P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{v}_{t+h}\right),$$

and $\widehat{V}_{\alpha}(R)$ is a consistent estimator of V_{α}. Under the null hypothesis $H_0:\lim_{T\to\infty}E\left[\widehat{\alpha}\left(R\right)\right]=0$ for all R,

$$\mathcal{R}_T^{\mathcal{W}}\Longrightarrow\sup_{\mu\in[\underline{\mu},\overline{\mu}]}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]^{\prime}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right], \tag{31}$$

and

$$\mathcal{A}_T^{\mathcal{W}}\Longrightarrow\int_{\underline{\mu}}^{\overline{\mu}}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]^{\prime}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]d\mu, \tag{32}$$

where $R=\left[\mu T\right]$, $\underline{R}=[\underline{\mu}T]$, $\overline{R}=\left[\overline{\mu}T\right]$, and $\mathcal{B}_p\left(\cdot\right)$ is a standard p-dimensional Brownian motion. The null hypothesis for the $\mathcal{R}_T^{\mathcal{W}}$ test is rejected when $\mathcal{R}_T^{\alpha}>k_{\alpha,p}^{\mathcal{R},\mathcal{W}}$ whereas the null hypothesis for the $\mathcal{A}_T^{\mathcal{W}}$ test is rejected when $\mathcal{A}_T^{\alpha}>k_{\alpha,p}^{\mathcal{A},\mathcal{W}}$. Simulated values of $\left(\alpha,k_{\alpha,p}^{\mathcal{R},\mathcal{W}}\right)$ and $\left(\alpha,k_{\alpha,p}^{\mathcal{A},\mathcal{W}}\right)$ for various values of $\underline{\mu}$, $\overline{\mu}=1-\underline{\mu}$ and p are reported in Table 3.[19]

[18] The traditional Wald test would be: $\widehat{\alpha}\left(R\right)^{\prime}\ \widehat{V}_{\alpha}^{-1}(R)\widehat{\alpha}\left(R\right)$.

[19] The 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 $P\left\{\sup_{\mu\in[\underline{\mu},\overline{\mu}]}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]^{\prime}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]>k_{\alpha}^{\mathcal{R},\mathcal{W}}\right\}=\alpha$, $P\left\{\int_{\underline{\mu}}^{\overline{\mu}}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]^{\prime}\left[\mathcal{B}_p\left(1\right)-\mathcal{B}_p\left(\mu\right)\right]d\mu>k_{\alpha}^{\mathcal{A},\mathcal{W}}\right\}=\alpha$.

A consistent estimate of the asymptotic variance of $\widehat{\alpha}(R)$, when parameter estimation error is not relevant, is: $\widehat{V}_{\alpha}(R) = P^{-1}S_{gg}^{-1}(R)\widehat{\Omega}_{R}S_{gg}^{-1}(R)$, where $S_{gg}^{-1} = \left(P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{g}_t'\right)^{-1}$ and $\widehat{\Omega}_R = \sum_{i=-q(P)+1}^{q(P)-1}(1-|i/q(P)|)P^{-1}\sum_{t=R}^{T}\widehat{g}_t\widehat{g}_{t-i}'\widehat{v}_{t+h}\widehat{v}_{t+h-i}$. See West and McCracken (1998) for alternative and more general variance estimators.

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

2.4 Robust Point Forecasts

The techniques discussed in this paper can also be used to construct robust confidence intervals for point forecasts. Let $\widehat{y}_{t,t+h}(R)$ denote the forecast of variable y_{t+h} made at time t with a direct forecast method, using data estimated over a window of size R. It is practically convenient to assume that the data are stationary, the forecast model is correctly specified, and the forecast errors are optimal and normally distributed with zero mean and variance σ^2 conditional on the regressors (cfr. Stock and Watson, 2003b, p. 451). In such cases, a $(1-\alpha)100\%$ forecast interval is given by $\widehat{y}_{t,t+h}(R) \pm z_{\alpha}\cdot\widehat{\sigma}_{y_{t+h}-\widehat{y}_{t,t+h}}(R)$ where z_{α} is the critical value for the standard normal distribution at the $100\alpha\%$ significance level (e.g. 1.96 for a 5% significance level), and $\widehat{\sigma}_{y_{t+h}-\widehat{y}_{t,t+h}}(R)$ is an estimator of the root mean squared forecast error; for example, one could use the square root of a HAC robust estimator of the variance of the pseudo-out of sample forecasts or the standard error of the regression. However, again such forecasts are obtained conditional upon the choice of the rolling window size.

We propose a robust point forecast obtained by averaging the forecasts obtained over all possible window sizes, that is:

$$\widehat{y}_{t,t+h}^{\mathcal{A}} = \frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}}\widehat{y}_{t,t+h}(R)$$

Assuming that parameter estimation error is asymptotically negligible, the estimate of the root mean squared forecast error will be consistent no matter which window size it is constructed with. We propose the researcher to use the estimate of the root mean squared forecast error based on the largest window size, $\widehat{\sigma}_{y_{t+h}-\widehat{y}_{t,t+h}}\left(\overline{R}\right)$. Then, our proposed $\alpha - 100\%$ forecast interval is: $\widehat{y}_{t,t+h}^{\mathcal{A}} \pm z_{\alpha}\cdot\widehat{\sigma}_{y_{t+h}-\widehat{y}_{t,t+h}}\left(\overline{R}\right)$.

3 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. For the nested models comparison, we consider two Data Generating Processes (DGPs). Let the first (labeled "DGP1") be:

$$y_{t+1} = \alpha_t x_t + \gamma_t z_t + \varepsilon_{t+1},\ t = 1, ..., T,$$

where x_t, ε_{t+1} and z_t are all i.i.d. standard normals and independent of each other. We compare the following two nested models' forecasts for y_t:

$$\begin{aligned} \text{Model 1 forecast} &: \quad \widehat{\alpha}_t x_t \\ \text{Model 2 forecast} &: \quad \widehat{\theta}_t^{\prime} w_t, \end{aligned} \tag{33}$$

where $w_t = [x_t, z_t]'$, and both models' parameters are estimated by OLS in rolling windows of size R: $\widehat{\alpha}_t(R) = \left(\Sigma_{j=t-R}^{t-1} x_j y_{j+1}\right)\left(\Sigma_{j=t-R}^{t-1} x_j^2\right)^{-1}$ and $\widehat{\theta}_t(R) = \left(\Sigma_{j=t-R}^{t-1} w_j w_j'\right)^{-1}\left(\Sigma_{j=t-R}^{t-1} w_j y_{j+1}\right)$ for $t = R, ..., T$.

We also consider a second and more realistic DGP (labeled "DGP2") that follows Clark and McCracken (2005a) and Pesaran and Timmermann (2007). Let

$$\begin{pmatrix} y_{t+1} \\ x_{t+1} \end{pmatrix} = \begin{pmatrix} 0.3 & 0.5d_t \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} + \begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix},\ t = 1, ..., T-1,$$

where $y_0 = x_0 = 0$, and $u_{y,t}$ and $u_{x,t}$ are both i.i.d. standard normal and independent of each other. $d = 1$ if $t \leq \tau$ and 0 otherwise. By varying τ, we evaluate cases ranging from when there is no break ($\tau = 0$) to cases when there is a break ($0 < \tau < T$).

We compare the following two nested models' forecasts for y_{t+1}:

$$\begin{aligned} \text{Model 1 forecast} &: \quad \widehat{\gamma}_0 + \widehat{\gamma}_1 y_t \\ \text{Model 2 forecast} &: \quad \widehat{\gamma}_0 + \widehat{\gamma}_1 y_t + \widehat{\gamma}_2 x_t, \end{aligned} \tag{34}$$

and both models's parameters are estimated by OLS in rolling windows of size R.

For the non-nested models' comparison, we consider a third DGP (labeled "DGP3"):

$$y_{t+1} = \delta_t x_t + \beta_t z_t + \varepsilon_t,\ t = 1, ..., T,$$

where x_t and z_t are standard normals independent of each other. We compare the following two non-nested models' forecasts for y_t:

$$\begin{aligned} \text{Model 1 forecast} &: \quad \widehat{\delta}_t x_t \\ \text{Model 2 forecast} &: \quad \widehat{\beta}_t z_t, \end{aligned} \tag{35}$$

and both models's parameters are estimated by OLS in rolling windows of size R.

3.1 Size properties

The size properties of our test procedures in small samples are evaluated in a series of Monte Carlo experiments. The null hypothesis discussed in Propositions (2) and (3) holds in DGP1 if $\alpha_t = 1$, $\gamma_t = 0$ for $t = 1, 2, ...T$, and the same null hypothesis is satisfied in DGP2 if $d_t = 0$ for $t = 1, 2, ...T$.[20] The null hypothesis discussed in Proposition (1) holds in DGP3 holds if $\delta_t = \beta_t = 1$ for $t = 1, 2, ...T$. We investigate sample sizes where $T = 50$, 100, 200 and 500. In all experiments, we set $\underline{\mu} = 0.15$. The number of Monte Carlo replications is 5,000.

We report empirical rejection probabilities of the tests $\mathcal{R}_T$ and $\mathcal{A}_T$ at the 10%, 5%, and 1% nominal levels. Table 4(a) reports results for the nested models' comparison for DGP1; results for DGP2 are reported in Table 4(b) whereas Table 4(c) reports results for the non-nested models' comparison (DGP3). All tables show that each test has good size properties, although they may slightly under/over-reject for small sample sizes.

INSERT TABLES 4(a,b,c)

3.2 Power properties

The scope of this sub-section is three-fold. First, we evaluate the power properties of our proposed procedure to departures from the null hypothesis in small samples. Second, we show that the existing methods, which rely on an “ad-hoc” window size choice, may have no power at all to detect predictive ability. Third, we demonstrate that existing methods are subject to data mining if they are applied to many window sizes without correcting the appropriate critical values.

For the nested models comparison, in DGP1 we let $\alpha_t = 1$, $\gamma_t = 1 \cdot 1\left(t \leq \tau\right) + 0 \cdot 1\left(t > \tau\right)$, for τ as in Table 5(a). For DGP2, we let $d_t = 1 \cdot 1\left(t \leq \tau\right) + 0 \cdot 1\left(t > \tau\right)$, for τ as in Table 5(b). For the non-nested models' comparison, we let the DGP3 be such that $\delta_t = 1$, $\beta_t = 0 \cdot 1\left(t \leq \tau\right) + 1 \cdot 1\left(t > \tau\right)$, for τ as in Table 5(c).

The column labeled "$\mathcal{R}_T$ test" reports empirical rejection rates for (2) and the column labeled "$\mathcal{A}_T$ test" reports empirical rejection rates for (3). In addition, in Tables 5(a,b) the columns labeled "Fixed R" report empirical rejection rates for the Clark and West's (2007) test implemented with a specific value of R, for which the one-sided critical value is 1.645 for

[20] Note that this corresponds to setting $\tau = 0$.

a 5% nominal level test. The "Fixed R" results 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. Similarly, in Table 5(c), the columns labeled "Fixed R" report rejection rates for the Diebold and Mariano's (1995), West's (1996) and McCracken's (2001) test implemented with a specific value of R, for which the two-sided critical value is 1.96 for a 5% nominal level test.[21] Finally, the columns labeled "Data Mining" report empirical rejection rates incurred by a researcher who is searching across all values of $R \in [0.15T, 0.85T]$ but using the Clark and West (2007) test's critical value in Tables 5(a,b) and the Diebold and Mariano (1995), West (1996) and McCracken (2001) test's critical value in Table 5(c) (that is, without taking into account the search procedure). The latter columns show how data snooping distorts inference when it is not properly taken into account. The empirical rejection rates are shown for various values of the break point, τ. The first row in each table reports results for size.

Table 5(a) shows that all tests have approximately the correct size except the "Data mining" procedure, which has size distortions, and leads to too many rejections. 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. The remaining rows show the power of the tests. They demonstrate that, in general, in the presence of a structural break our proposed tests have better power than the tests based on a specific rolling window. The latter may have negligible power for some combinations of the location of the break (τ) and the size of the estimation window (R). Similar results hold for Tables 5(b) and 5(c).

INSERT TABLES 5(a,b,c)

4 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. The Meese

[21] The three tests differ because a different estimate of the variance is suggested. Diebold and Mariano (1995) suggest no parameter estimation correction, and West (1996) proposed an estimate that includes parameter estimation correction. McCracken (2001) extended West's (1996) results to rolling windows and more general loss functions. In our simulations, these tests are the same since parameter estimation uncertainty is asymptotically irrelevant.

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 (2009) evaluated the consequences of instabilities in the forecasting performance of the models over time; and 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.[22] 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,

$$s_{t+1} - s_t = \alpha + \beta z_t + \varepsilon_{t+1} \tag{36}$$

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,

$$f_t^{UIRP} = (i_t - i_t^*) + s_t, \tag{37}$$

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

[22] The exchange rate is defined as the domestic price of foreign currency.

The second model we consider is a model with Taylor rule fundamentals, as in Molodtsova and Papell (2007) 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,[23] y_t^{gap*} denote the output gap in the foreign country. Since the difference in the Taylor rule of the home and foreign countries implies $i_t - i_t^* = \delta\left(\pi_t - \pi_t^*\right) + \gamma\left(y_t^{gap} - y_t^{gap*}\right)$, we have that the latter determines the long run equilibrium level of the nominal exchange rate:

$$f_t^{TAYLOR} = \delta\left(\pi_t - \pi_t^*\right) + \gamma\left(y_t^{gap} - y_t^{gap*}\right) + s_t. \tag{38}$$

The benchmark model, against which the forecasts of both models (37 and 38) are evaluated, is the random walk, according to which the exchange rate changes are forecasted to be zero.[24]

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.[25] 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 Proposition 2 for comparability with Molodtsova and Papell (2009), who use the Clark and West's (2007) test. 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 in Figures 1 and 2, offering 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 6, and Figures 1 and 2. The column labeled "Fixed R Test" in Table 6 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 (2007). According to the "Fixed R test", the Taylor model is significantly outperforming a random walk for Canada

[23] 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.

[24] 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).

[25] Data on interest rates were incomplete for Portugal and the Netherlands, so we do not report UIRP results for these countries.

at 5% and for Japan at 10%, whereas the UIRP model outperforms the random walk for Canada at 5% and both Japan and the U.K. at the 10% 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 West's (2007) test statistic for the window sizes that we consider. In particular, the predictive ability of the economic models tend to show up at smaller window sizes, as the figures show.

INSERT TABLE 6 AND FIGURES 1 AND 2

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

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

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

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

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

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

[11] Clark, T. 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.

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

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

[14] 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.

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

[16] 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).

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

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

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

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

[21] McCracken, M. (2000), “Robust Out-of-Sample Inference”, *Journal of Econometrics,* 99, 195-223.

[22] 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.

[23] 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.

[24] 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, pp. 81–111.

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

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

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

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

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

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

[31] Pesaran, M.H. and A. Timmermann (2007), "Selection of estimation window in the presence of breaks", *Journal of Econometrics* 137(1), 134-161.

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

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

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

[35] Rossi, B. and T. Sekhposyan (2010), “Understanding Models’ Forecasting Performance”, *mimeo*, Duke University.

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

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

[38] 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.

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

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

6 Appendix A. Proofs

Proof of Proposition 1. For the $\mathcal{R}_T$ test, note that from Assumption 1:

$$\sigma^{-1}T^{-1/2}\sum_{t=R}^{T}\Delta L_{t+h}(\widehat{\theta}_{t,R},\widehat{\gamma}_{t,R})\Rightarrow\mathcal{B}\left(\mu\right).$$

Under H_0, $\widehat{\sigma}$ is a consistent HAC estimator of σ (Newey and West, 1987). Since the absolute value and the Sup(.) function are continuous functions, the Theorem follows from the Continuous Mapping Theorem. The proof is similar for the $\mathcal{A}_T$ test. ■

Proof of Proposition 2. The proof is similar to that of Proposition 1, and it is therefore omitted. ■

Proof of Proposition 3. From Assumption 3 and Lemma A6 in Clark and McCracken (2000), under H_0 we have

$$\Delta L_T^{\mathcal{E}}\left(R\right)\underset{d}{\rightarrow}\mu^{-1}\int_{\mu}^{1}\left[\mathcal{B}_k\left(s\right)-\mathcal{B}_k\left(s-\mu\right)\right]'d\mathcal{B}_k\left(s\right).$$

The result follows from the Continuous Mapping Theorem. ■

Proof of Proposition 4. For the $\mathcal{R}_T^{\mathcal{W}}$ test, note that under Assumption 4:

$$T^{-1/2}\Omega^{-1/2}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})\Rightarrow\mathcal{B}_p\left(\mu\right)$$

Then, $\mathcal{L}_T\left(R\right)'\mathcal{L}_T\left(R\right)=\left[T^{-1/2}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})\right]'\Omega^{-1}\left[T^{-1/2}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}(\widehat{\theta}_{t,R})\right]\Rightarrow\mathcal{B}_p\left(\mu\right)'\mathcal{B}_p\left(\mu\right).$ Under H_0, $\widehat{\Omega}$ is a consistent HAC estimator of Ω (Newey and West, 1987). The Sup(.) function is a continuous function and the Theorem follows from the Continuous Mapping Theorem. The proof is similar for the $\mathcal{A}_T^{\mathcal{W}}$ test. ■

Proof of Proposition 5. The proof is similar to that of Proposition 4, and it is therefore omitted. ■

7 Appendix B: The Recursive Case

We consider the recursive scheme in which the researcher estimates forecasting models using first t observations and make h-steps ahead forecasts for $t = R, R+1, ..., T$. Let $\hat{\theta}_t$ and $\hat{\gamma}_t$ denote the recursive estimates of θ^* and γ^* based on the first t observations, and let $\{\Delta L_{t+h}(\hat{\theta}_t, \hat{\gamma}_t)\}_{t=R}^{T}$ denote a sequence of loss differences of forecasting models 1 and 2. Below $\mathcal{A}_T$, $\mathcal{A}_T^{\mathcal{E}}$, $\Delta L_T(R)$, $\Delta L_T^{adj}(R)$, $\Delta L_T^{\mathcal{E}}(R)$, $\mathcal{R}_T$, $\mathcal{R}_T^{\mathcal{E}}$, $\mathcal{W}_T(R)$ and $\mathcal{W}_T^{(r)}(R)$ are the same as those in those in Section 2 except that rolling estimates $\hat{\theta}_{t,R}$ and $\hat{\gamma}_{t,R}$ replaced by recursive estimates $\hat{\theta}_t$ and $\hat{\gamma}_t$, respectively.

7.1 Non-Nested Model Comparisons

Assumption 1': (a) The partial sum $T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\{\Delta L_{t+h}(\hat{\theta}_t, \hat{\gamma}_t) - E[\Delta L_{t+h}(\theta^*, \gamma^*)]\}$ *obeys a functional central limit theorem:*

$$\sigma_R^{-1}T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\left\{\Delta L_{t+h}(\hat{\theta}_t, \hat{\gamma}_t) - E[\Delta L_{t+h}(\theta^*, \gamma^*)]\right\} \Rightarrow \mathcal{B}(\cdot) - \mathcal{B}(\mu), \tag{39}$$

where σ^2 *is the long-run variance of loss differences; and (b)* $lim_{T,R\to\infty} R/T = \mu \in (0,1)$.

Proposition 6 (Out-of-sample robust test for non-nested models) *Suppose Assumption 1' holds. Under the null hypothesis* $H_0 : \lim_{T\to\infty} E\left[\Delta L_T^*(R)\right] = 0$ *for all* R,

$$\mathcal{R}_T \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]} \left|\mathcal{B}(1) - \mathcal{B}(\mu)\right|, \tag{40}$$

and

$$\mathcal{A}_T \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left|\mathcal{B}(1) - \mathcal{B}(\mu)\right| d\mu, \tag{41}$$

where $\mathcal{B}(\cdot)$ *is a standard univariate Brownian motion.*

7.2 Nested Model Comparison

Assumption 2': (a) The partial sum $T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\left\{\Delta L_{t+h}^{adj}(\hat{\theta}_t, \hat{\gamma}_t) - E[\Delta L_{t+h}^{adj}(\theta^*, \gamma^*)]\right\}$ *obeys a functional central limit theorem:*

$$\sigma^{-1}T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\left\{\Delta L_{t+h}^{adj}(\hat{\theta}_t, \hat{\gamma}_t) - E[\Delta L_{t+h}^{adj}(\theta^*, \gamma^*)]\right\} \Rightarrow \mathcal{B}(\cdot) - \mathcal{B}(\mu), \tag{42}$$

where σ^2 *is the long-run variance of loss differences; and (b)*$lim_{T,R\to\infty} R/T = \mu \in (0,1)$

Proposition 7 (Out-of-sample robust test for nested models) *Suppose Assumption 2' holds. Under the null hypothesis* $H_0 : \lim_{T\to\infty} E[\Delta L_T^{adj}(R)] = 0$ *for all* R,

$$\mathcal{R}_T \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]} \left[\mathcal{B}(1) - \mathcal{B}(\mu)\right], \tag{43}$$

and

$$\mathcal{A}_T \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left[\mathcal{B}(1) - \mathcal{B}(\mu)\right] d\mu, \tag{44}$$

where $R = [\mu T]$, $\underline{R} = [\underline{\mu}T]$, $\overline{R} = [\overline{\mu}T]$, *and* $\mathcal{B}(\cdot)$ *is a standard univariate Brownian motion.*

Assumption 3': (a) The parameter estimates $\hat{\theta}_t$ *satisfies* $\hat{\theta}_t - \theta^* = B_1(t) H_1(t)$ *where* $B_1(t) H_1(t) = \left(R^{-1}\sum_{j=1}^{t} q_{1,j}\right)^{-1}\left(R^{-1}\sum_{j=1}^{t} h_{1,j}\right)$, *and similarly for* $\hat{\gamma}_t - \gamma^* = B_2(t) H_2(t)$; *(b) Let* $U_t = \left[u_t, x_{2,t}' - Ex_{2,t}', h_{2,t}', vec\left(h_{2,t}h_{2,t}' - Eh_{2,t}h_{2,t}'\right)', vec\left(q_{2,t} - Eq_{2,t}\right)'\right]'$. *Then* $EU_t = 0$; $Eq_{2,t} < \infty$ *is p.d.; for some* $r > 4$, U_t *is uniformly* L^r *bounded; for all* t, $Eu_t^2 = \sigma^2 < \infty$; *for some* $r > d > 2$, U_t *is strong mixing with coefficients of size* $-rd/(r-d)$; *letting* $\widetilde{U}_t$ *denote the vector of non-redundant elements of* U_t, $lim_{T\to\infty}T^{-1}E\left(\sum_{t=1}^{T}\widetilde{U}_t\right)\left(\sum_{t=1}^{T}\widetilde{U}_t\right)' = \Omega < \infty$ *is p.d.; (c)* $E\left(h_{2,t}h_{2,t}'\right) = \sigma^2 Eq_{2,t}$ *and* $E\left(h_{2,t}|h_{2,t-j}, q_{2,t-j,}j = 1, 2, ...\right) = 0$ *; and (d)* $lim_{T,R\to\infty}R/T = \mu \in (0,1)$.

Proposition 8 (Out-of-sample robust test for nested models II) *Suppose Assumption 3' holds. Under the null hypothesis* $H_0 : \lim_{T\to\infty} E[\Delta L_T^{\mathcal{E}}(R)] = 0$ *for all* R,

$$\mathcal{R}_T^{\mathcal{E}} \Longrightarrow \sup_{\mu\in[\underline{\mu},\overline{\mu}]} \int_{\mu}^{1} s^{-1}\mathcal{B}_k(s)' \, d\mathcal{B}_k(s), \tag{45}$$

and

$$\mathcal{A}_T^{\mathcal{E}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left\{\int_{\mu}^{1} s^{-1}\mathcal{B}_k(s)' \, d\mathcal{B}_k(s)\right\} d\mu, \tag{46}$$

where $\underline{R} = [\underline{\mu}T]$, $\overline{R} = [\overline{\mu}T]$, *and* $\mathcal{B}_k(\cdot)$ *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.*

Table A.1 provides the critical values for the $\mathcal{R}_T^{\mathcal{E}}$ and $\mathcal{A}_T^{\mathcal{E}}$ tests.

7.3 Regression-based tests of predictive ability

Assumption 4': (a) The partial sum $T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\left\{\mathcal{L}_{t+h}(\hat{\theta}_{t,R}) - E[\mathcal{L}_{t+h}(\theta^*)]\right\}$ *obeys a functional central limit theorem:*

$$\Omega^{-\frac{1}{2}}T^{-1/2}\sum_{t=[\mu T]}^{[\cdot T]}\left\{\mathcal{L}_{t+h}(\hat{\theta}_{t,R}) - E[\mathcal{L}_{t+h}(\theta^*)]\right\} \Rightarrow \mathcal{B}_p(\cdot) - \mathcal{B}_p(\mu) \tag{47}$$

where $B_p(\cdot)$ denotes the p-dimensional standard Brownian motion; (b) $lim_{T,R\to\infty} R/T = \mu \in (0,1)$.

Proposition 9 (Robust Regression-Based Tests) *Suppose Assumption 4' holds. Under the null hypothesis $H_0 : \lim_{T\to\infty} E\left(\mathcal{L}_{t+h}(\theta^*)\right) = 0$ for all R,*

$$\mathcal{R}_T^{\mathcal{W}} \Longrightarrow \sup_{\mu \in [\underline{\mu},\overline{\mu}]} \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right]' \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right], \tag{48}$$

and

$$\mathcal{A}_T^{\mathcal{W}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right]' \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right] d\mu, \tag{49}$$

where $R = [\mu T]$, $\underline{R} = [\underline{\mu} T]$, $\overline{R} = [\overline{\mu} T]$, and $\mathcal{B}_p(\cdot)$ is a standard p-dimensional Brownian motion.

Proposition 10 (Robust Regression-Based Tests II) *Suppose Assumption 4' holds and $E\left(g_t g_t'\right)$ is full rank. Under the null hypothesis $H_0 : E\left[\widehat{\alpha}\left(R\right)\right] = 0$ for all R*

$$\mathcal{R}_T^{\mathcal{W}} \Longrightarrow \sup_{\mu \in [\underline{\mu},\overline{\mu}]} \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right]' \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right], \tag{50}$$

and

$$\mathcal{A}_T^{\mathcal{W}} \Longrightarrow \int_{\underline{\mu}}^{\overline{\mu}} \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right]' \left[\mathcal{B}_p\left(1\right) - \mathcal{B}_p\left(\mu\right)\right] d\mu, \tag{51}$$

where $R = [\mu T]$, $\underline{R} = [\underline{\mu} T]$, $\overline{R} = [\overline{\mu} T]$, and $\mathcal{B}_p(\cdot)$ is a standard p-dimensional Brownian motion.

8 Tables and Figures

Table 1. Critical Values for Non-Nested Model Comparisons

	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
$\underline{\mu}$	10%	5%	1%	10%	5%	1%
0.15	1.7843	2.0456	2.5539	1.0194	1.2132	1.5888
0.20	1.7260	1.9663	2.4670	1.0392	1.2331	1.6221
0.25	1.6713	1.9158	2.4042	1.0639	1.2668	1.6545
0.30	1.6173	1.8521	2.3418	1.0865	1.2911	1.7024
0.35	1.5489	1.7703	2.2368	1.1041	1.3161	1.7201

Notes to Table 1. The Critical Values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications.

Table 2(a). Critical Values for Nested Model Comparisons Using Clark and West (2007)

	$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
$\underline{\mu}$	10%	5%	1%	10%	5%	1%
0.15	1.4929	1.7849	2.3444	0.7969	1.0147	1.4374
0.20	1.4449	1.7226	2.2583	0.8112	1.0363	1.4606
0.25	1.3955	1.6731	2.2190	0.8244	1.0651	1.5005
0.30	1.3487	1.6169	2.1386	0.8476	1.0831	1.5385
0.35	1.2939	1.5494	2.0436	0.8619	1.1043	1.5588

Notes to Table 2(a). The Critical Values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications.

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW

Panel A. 10% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		3.8443	3.2051	2.7401	2.3870	2.0917	1.0913	1.0621	1.0825	1.0898	1.0994
2		5.5136	4.5857	3.9968	3.4806	3.0744	1.6350	1.6290	1.6461	1.6685	1.6562
3		6.6595	5.6350	4.8969	4.2868	3.8388	2.0266	2.0149	2.0544	2.0848	2.1326
4		7.6973	6.5229	5.6131	4.9877	4.4630	2.3621	2.3665	2.3424	2.3971	2.4885
5		8.6283	7.2211	6.3395	5.6046	5.0276	2.6276	2.6094	2.6975	2.6700	2.7890
6		9.3927	8.0222	6.9461	6.1068	5.4293	2.8703	2.9229	2.8961	2.9423	2.9936
7		10.2100	8.6033	7.4511	6.5518	5.8925	3.1087	3.0821	3.1732	3.1663	3.2575
8		10.9141	9.1959	8.0006	7.0794	6.2935	3.3849	3.3038	3.3747	3.4183	3.4947
9		11.5085	9.7622	8.4207	7.4859	6.6801	3.5824	3.5587	3.6026	3.6250	3.6966
10		12.1040	10.2965	8.9722	7.9025	7.0086	3.7444	3.7548	3.8395	3.8266	3.8641
11		12.6332	10.8432	9.3603	8.2559	7.3431	3.8839	3.9848	3.9521	4.0226	4.0849
12		13.2473	11.2027	9.8130	8.6287	7.6146	4.0391	4.0675	4.1412	4.1952	4.1868
13		13.8136	11.6817	10.1103	8.9722	8.0089	4.2627	4.2586	4.2692	4.3033	4.3963
14		14.1817	12.1014	10.5226	9.3049	8.3163	4.3900	4.4456	4.4273	4.5474	4.5813

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW

Panel A. 5% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		5.1436	4.2982	3.7649	3.3010	2.9402	1.7635	1.7335	1.7737	1.7964	1.7998
2		7.1284	5.9983	5.2541	4.6326	4.1224	2.4879	2.4475	2.4953	2.5174	2.5074
3		8.4892	7.1863	6.3640	5.6474	5.0395	2.9559	2.9653	3.0244	3.0510	3.1128
4		9.7745	8.3132	7.1991	6.4048	5.7775	3.3900	3.4071	3.3971	3.4840	3.5546
5		10.8235	9.1526	8.0207	7.1808	6.4377	3.7427	3.7055	3.8169	3.8627	3.9389
6		11.6763	10.0692	8.7422	7.7794	6.9733	4.0485	4.0943	4.0928	4.1238	4.2508
7		12.7226	10.7869	9.4653	8.2669	7.4864	4.3890	4.3478	4.4171	4.4408	4.5547
8		13.4939	11.5170	9.9787	8.9054	7.9574	4.6356	4.6621	4.6983	4.7957	4.8064
9		14.3192	12.1188	10.5791	9.4202	8.4937	4.9058	4.9153	4.9987	5.0139	5.1388
10		14.9491	12.7733	11.2190	9.9213	8.8382	5.1336	5.1859	5.2762	5.2946	5.3531
11		15.6930	13.4913	11.6504	10.3327	9.2877	5.3506	5.4240	5.4829	5.4919	5.5748
12		16.2954	13.8349	12.2591	10.7844	9.5457	5.5861	5.5867	5.6622	5.7151	5.7248
13		17.0991	14.5326	12.5030	11.1212	10.0114	5.7750	5.8441	5.8337	5.9103	6.0124
14		17.4946	14.9149	13.0349	11.5651	10.3584	5.9958	6.0267	6.0220	6.1882	6.2575

Table 2(b). Critical Values for Nested Model Comparisons Using ENCNEW

Panel A. 1% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{E}}$ test					$\mathcal{A}_T^{\mathcal{E}}$ test				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		7.9243	6.8435	6.2532	5.7046	5.1323	3.3873	3.3839	3.4538	3.5434	3.5885
2		10.6078	9.1073	8.1681	7.2456	6.6399	4.5099	4.4008	4.4883	4.4810	4.5776
3		12.3911	10.7002	9.6115	8.6454	7.7958	5.0938	5.0960	5.3018	5.3210	5.2876
4		13.9680	12.1513	10.5579	9.5567	8.7154	5.6628	5.8212	5.6318	5.8305	5.8822
5		15.5409	13.3539	11.7835	10.5474	9.5755	6.1676	6.1541	6.2896	6.3981	6.4738
6		16.7209	14.5039	12.6194	11.3325	10.3700	6.7097	6.6566	6.7122	6.7471	6.9580
7		18.0511	15.3139	13.6682	12.1613	10.9275	7.0585	6.9986	7.1478	7.1808	7.2750
8		18.8899	16.3693	14.2198	12.9446	11.4624	7.3854	7.5575	7.4173	7.6441	7.6924
9		20.1286	17.1119	15.0996	13.6190	12.3904	7.8039	7.8023	7.8329	8.0351	8.2174
10		21.0123	17.9317	15.9820	14.1687	13.0504	8.0456	8.1048	8.3548	8.3590	8.5415
11		22.1865	18.7922	16.5889	14.8316	13.3213	8.4460	8.4763	8.5434	8.8163	8.7738
12		22.6814	19.3941	17.3299	15.3705	13.7038	8.7419	8.8274	8.9273	8.8940	8.9904
13		23.7040	20.2035	17.6548	15.7804	14.0921	9.2106	9.2550	9.0222	9.2913	9.3220
14		24.1862	20.9349	18.2787	16.2610	14.5866	9.2207	9.2680	9.4063	9.6465	9.6627

Notes to Table 2(b). The Critical Values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications.

Table 3. Critical Values for Regression-Based Forecast Tests

Panel A. 10% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
p	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		0.925	0.873	0.820	0.768	0.716	0.550	0.552	0.553	0.554	0.555
2		1.806	1.703	1.600	1.496	1.393	1.071	1.073	1.074	1.075	1.077
3		2.680	2.526	2.372	2.217	2.062	1.587	1.588	1.592	1.593	1.594
4		3.549	3.345	3.142	2.936	2.730	2.100	2.103	2.106	2.107	2.108
5		4.418	4.162	3.908	3.653	3.396	2.612	2.615	2.617	2.619	2.622
6		5.283	4.977	4.674	4.365	4.059	3.123	3.126	3.128	3.129	3.133
7		6.147	5.791	5.435	5.079	4.723	3.632	3.635	3.638	3.640	3.644
8		7.014	6.605	6.198	5.791	5.385	4.144	4.145	4.148	4.1511	4.153
9		7.874	7.419	6.960	6.504	6.049	4.652	4.654	4.657	4.660	4.664
10		8.738	8.232	7.724	7.218	6.706	5.159	5.163	5.167	5.170	5.172
11		9.598	9.045	8.483	7.925	7.368	5.666	5.671	5.674	5.678	5.681
12		10.458	9.853	9.242	8.636	8.028	6.176	6.178	6.180	6.186	6.188
13		11.319	10.664	10.003	9.345	8.686	6.682	6.686	6.689	6.694	6.696
14		12.182	11.472	10.762	10.054	9.344	7.189	7.194	7.195	7.201	7.204
15		13.037	12.279	11.525	10.766	9.999	7.691	7.698	7.703	7.707	7.708

Table 3. Critical Values for Regression-Based Forecast Tests

Panel B. 5% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
p	$\underline{\mu} =$	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		0.947	0.895	0.842	0.790	0.736	0.565	0.568	0.569	0.570	0.571
2		1.836	1.734	1.629	1.524	1.421	1.092	1.095	1.096	1.097	1.101
3		2.718	2.563	2.407	2.252	2.095	1.613	1.614	1.618	1.620	1.622
4		3.594	3.388	3.183	2.976	2.768	2.131	2.132	2.137	2.138	2.140
5		4.465	4.208	3.952	3.697	3.439	2.646	2.648	2.652	2.655	2.656
6		5.336	5.028	4.722	4.415	4.104	3.159	3.162	3.165	3.167	3.171
7		6.204	5.845	5.491	5.132	4.772	3.671	3.674	3.678	3.682	3.686
8		7.076	6.664	6.255	5.846	5.440	4.186	4.188	4.191	4.194	4.199
9		7.939	7.482	7.022	6.563	6.103	4.696	4.697	4.702	4.706	4.712
10		8.805	8.299	7.789	7.280	6.764	5.204	5.209	5.216	5.220	5.221
11		9.668	9.112	8.549	7.988	7.429	5.713	5.720	5.724	5.729	5.733
12		10.535	9.924	9.313	8.702	8.095	6.227	6.230	6.232	6.242	6.243
13		11.397	10.740	10.076	9.418	8.756	6.734	6.740	6.744	6.749	6.755
14		12.262	11.552	10.838	10.128	9.414	7.242	7.249	7.250	7.260	7.261
15		13.119	12.358	11.604	10.839	10.071	7.748	7.755	7.763	7.766	7.769

Table 3. Critical Values for Regression-Based Forecast Tests

Panel C. 1% Nominal Significance Level

		$\mathcal{R}_T^{\mathcal{W}}$ test					$\mathcal{A}_T^{\mathcal{W}}$ test				
p	$\underline{\mu} =$	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		0.991	0.938	0.882	0.830	0.775	0.596	0.598	0.600	0.603	0.604
2		1.896	1.791	1.685	1.579	1.473	1.133	1.138	1.139	1.143	1.145
3		2.788	2.631	2.475	2.316	2.159	1.661	1.665	1.671	1.672	1.674
4		3.675	3.468	3.261	3.051	2.842	2.185	2.192	2.195	2.198	2.201
5		4.558	4.299	4.039	3.782	3.521	2.707	2.712	2.715	2.721	2.726
6		5.436	5.128	4.817	4.504	4.188	3.227	3.231	3.237	3.240	3.243
7		6.311	5.951	5.597	5.234	4.869	3.744	3.746	3.759	3.762	3.768
8		7.191	6.778	6.365	5.953	5.545	4.265	4.265	4.272	4.278	4.286
9		8.060	7.605	7.140	6.673	6.211	4.777	4.786	4.789	4.792	4.800
10		8.934	8.419	7.906	7.394	6.879	5.290	5.300	5.308	5.310	5.317
11		9.802	9.236	8.673	8.113	7.548	5.800	5.811	5.818	5.823	5.827
12		10.677	10.057	9.442	8.833	8.213	6.322	6.324	6.327	6.342	6.348
13		11.542	10.879	10.212	9.544	8.886	6.833	6.845	6.848	6.855	6.864
14		12.412	11.703	10.981	10.269	9.551	7.344	7.356	7.361	7.367	7.378
15		13.273	12.511	11.752	10.981	10.210	7.848	7.863	7.878	7.881	7.884

Notes to Table 3. The Critical Values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications.

Table 4(a). Size properties – DGP1

T	$\alpha =$	0.10	0.05	0.01	0.10	0.05	0.01
		$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
50		0.125	0.059	0.012	0.060	0.031	0.005
100		0.103	0.049	0.008	0.060	0.032	0.005
200		0.088	0.038	0.008	0.055	0.028	0.006
500		0.084	0.040	0.010	0.055	0.029	0.005
		$\mathcal{R}_T^{\mathcal{E}}$ test			$\mathcal{A}_T^{\mathcal{E}}$ test		
50		0.129	0.071	0.018	0.095	0.038	0.007
100		0.097	0.052	0.009	0.084	0.035	0.004
200		0.087	0.044	0.007	0.080	0.031	0.005
500		0.090	0.050	0.010	0.068	0.036	0.006

Table 4(b). Size properties – DGP2

T	$\alpha =$	0.10	0.05	0.01	0.10	0.05	0.01
		$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
50		0.170	0.089	0.018	0.057	0.030	0.005
100		0.120	0.065	0.012	0.054	0.030	0.006
200		0.101	0.045	0.008	0.049	0.025	0.005
500		0.085	0.045	0.007	0.056	0.025	0.006
		$\mathcal{R}_T^{\mathcal{E}}$ test			$\mathcal{A}_T^{\mathcal{E}}$ test		
50		0.213	0.131	0.045	0.102	0.045	0.007
100		0.135	0.073	0.019	0.086	0.035	0.006
200		0.105	0.049	0.012	0.071	0.032	0.004
500		0.086	0.048	0.008	0.074	0.031	0.003

Table 4(c). Size properties – DGP3

		$\mathcal{R}_T$ test			$\mathcal{A}_T$ test		
T	$\alpha =$	0.10	0.05	0.01	0.10	0.05	0.01
50		0.175	0.095	0.026	0.135	0.075	0.018
100		0.132	0.068	0.017	0.117	0.065	0.014
200		0.129	0.064	0.016	0.124	0.064	0.012
500		0.110	0.052	0.013	0.109	0.058	0.012

Table 5(a). Power Properties – DGP 1

τ	$\mathcal{R}_T$ test	$\mathcal{A}_T$ test	$\mathcal{R}_T^{\mathcal{E}}$ test	$\mathcal{A}_T^{\mathcal{E}}$ test	Fixed R					Data Mining
					$R = 20$	$R = 40$	$R = 60$	$R = 70$	$R = 80$	
0	0.05	0.03	0.05	0.03	0.04	0.03	0.03	0.04	0.04	0.20
10	0.06	0.02	0.07	0.04	0.04	0.03	0.04	0.04	0.04	0.26
30	0.52	0.03	0.87	0.35	0.42	0.04	0.05	0.05	0.06	0.77
50	0.97	0.30	0.99	0.97	0.97	0.38	0.05	0.05	0.06	0.99
70	1	0.92	1	1	1	0.97	0.39	0.06	0.06	1
90	1	1	1	1	1	1	0.98	0.87	0.50	1
100	1	1	1	1	1	1	1	1	0.96	1

Note: The table refers to nested models comparisons. $T = 100$, $\alpha = 0.05$, τ denotes the time of the structural break ($\tau = 0$ corresponds to the no break case, and $\delta_t = \beta_t = 1$).

Table 5(b). Power Properties – DGP 2

τ	$\mathcal{R}_T$ test	$\mathcal{A}_T$ test	$\mathcal{R}_T^{\mathcal{E}}$ test	$\mathcal{A}_T^{\mathcal{E}}$ test	Fixed R				Data Mining
					$R = 40$	$R = 80$	$R = 120$	$R = 160$	
200	0.05	0.03	0.05	0.03	0.03	0.03	0.03	0.04	0.17
175	0.29	0.23	0.28	0.20	0.23	0.21	0.26	0.37	0.65
150	0.78	0.71	0.78	0.69	0.73	0.65	0.66	0.74	0.94
125	0.96	0.94	0.96	0.83	0.94	0.92	0.87	0.89	0.99

Note: The table refers to nested models comparisons. $T = 200$, $\alpha = 0.05$, τ denotes the time of the structural break ($\tau = 200$ corresponds to the no break case).

Table 5(c). Power Properties – DGP3

τ	$\mathcal{R}_T$ Test	$\mathcal{A}_T$ Test	Fixed R Test				Data Mining
			$R = 20$	$R = 40$	$R = 60$	$R = 70$	
0	0.07	0.06	0.06	0.06	0.05	0.06	0.28
20	0.07	0.06	0.06	0.06	0.05	0.06	0.28
40	0.18	0.13	0.12	0.09	0.07	0.07	0.39
60	0.47	0.37	0.39	0.24	0.12	0.10	0.65
80	0.87	0.80	0.82	0.62	0.31	0.18	0.94
100	1	0.99	0.99	0.89	0.61	0.39	1

Note: The table refers to non-nested models comparisons. $T = 100$, $\alpha = 0.05$, τ denotes the time of the structural break ($\tau = 0$ indicates that there is no break).

Table 6. Empirical Results

	$\mathcal{R}_T$ Test		$\mathcal{A}_T$ Test		Fixed R Test	
	UIRP	Taylor	UIRP	Taylor	UIRP	Taylor
Japan	2.03**	1.35	0.96*	0.38	1.55*	1.47
Canada	2.09**	2.19**	0.53	1.19**	2.04**	2.43**
Switzerland	2.33**	- -	0.81*	- -	0.96	- -
U.K.	2.77**	0.68	0.36	-0.10	1.38*	0.54
France	0.74	2.49**	-0.51	0.17	-0.96	0.42
Germany	2.22**	1.40	0.56	0.10	0.85	-0.14
Italy	2.03**	2.95**	0.04	0.16	0.49	1.08
Sweden	2.42**	2.35**	-0.42	0.76	-1.59	0.99
The Netherlands	- -	1.88**	- -	-0.27	- -	-0.37
Portugal	- -	4.14**	- -	1.14**	- -	-0.04

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 Fixed R Test, we implemented a Clark and West (2007) test using $R = 120$; one-sided critical values at 5% and 10% significance values are 1.645 and 1.282.

Table A.1. 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				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		2.1827	1.9264	1.7879	1.4260	1.5916	0.8948	1.0787	0.9592	0.8253	1.0099
2		3.1801	2.9478	2.6594	2.3078	2.1908	1.3649	1.4239	1.5652	1.3065	1.2920
3		3.9146	3.6633	3.4011	3.0615	2.6588	1.7754	1.7590	1.7001	1.8438	1.6944
4		4.7417	4.2915	3.8623	3.8250	3.3252	2.0285	2.1314	2.0289	2.1751	1.9485
5		5.0480	4.6996	4.1217	4.2087	3.7308	2.0493	2.4242	2.1420	2.4664	2.3779
6		5.6378	5.0717	4.6734	4.4714	3.9274	2.3393	2.2415	2.4451	2.5173	2.4447
7		5.8468	5.8820	4.7711	5.0149	4.3855	2.4662	2.8280	2.5254	2.9126	2.6096
8		6.2322	6.1828	5.1969	4.6258	4.4351	2.8001	3.1090	2.6810	2.5876	2.8219
9		6.8219	6.4243	5.7112	5.5296	5.1766	3.1421	2.8892	2.8348	3.0700	3.1410
10		6.7028	6.2789	6.2257	5.6198	5.3043	2.7940	3.0528	3.1299	3.2162	3.2613
11		7.3123	6.6593	6.6319	6.0154	5.7181	3.0127	2.9701	3.3804	3.2990	3.6762
12		7.8022	7.4812	6.6942	6.1125	5.6641	3.5317	3.3753	3.3873	3.4190	3.6107
13		8.0998	7.3365	6.6997	6.5756	5.7137	3.5906	3.3773	3.5418	3.5555	3.5109
14		8.2287	7.3384	7.2055	6.4391	6.6329	3.2382	3.5441	3.6074	3.6762	4.3744

Table A.1. 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				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		3.0078	2.7519	2.5599	1.9907	2.2563	1.4955	1.5039	1.6247	1.2337	1.5222
2		4.2555	4.0880	3.5536	3.4613	3.1296	2.1339	1.9985	2.1779	2.0319	2.0155
3		5.0577	5.0933	4.8163	3.9635	3.5828	2.3919	2.5030	2.5470	2.5376	2.2816
4		6.1064	5.6270	4.9304	4.8950	4.0866	2.9668	3.0964	2.9057	2.9622	2.9778
5		6.3340	6.1398	5.3206	5.3653	4.7878	2.9717	3.0795	3.0435	3.1488	3.2085
6		7.4112	6.8839	6.2701	6.0270	4.8252	3.4424	3.2665	3.6043	3.2818	3.4061
7		7.7526	7.1460	6.0374	6.6942	5.5520	3.3073	3.7196	3.4741	4.0366	3.7042
8		7.9398	7.9244	6.6278	5.9797	5.8828	3.6604	4.0539	3.9295	3.6180	3.8040
9		9.0941	8.2817	7.3482	7.2577	7.0083	4.4619	4.1143	4.1880	4.3008	4.4988
10		8.7240	8.3543	8.1949	7.1712	6.9438	4.0097	4.2171	4.1991	4.4541	4.7212
11		9.4989	8.2225	8.5984	7.5053	7.3425	4.2346	3.9172	4.8147	4.6562	4.8656
12		9.6488	9.0317	9.1087	7.7116	7.4956	4.8887	4.5745	4.7542	4.5030	4.6967
13		10.8272	9.3142	8.4367	8.2197	7.1382	4.6932	4.5081	4.8684	4.8802	4.6407
14		10.3130	9.3451	9.2183	8.1447	8.5050	4.6401	4.6586	5.0358	5.1234	6.0534

Table A.1. 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				
k	$\underline{\mu}$=	0.15	0.20	0.25	0.30	0.35	0.15	0.20	0.25	0.30	0.35
1		5.5973	5.0777	4.3923	4.0048	4.5139	2.5440	2.6105	2.7804	2.4437	3.3912
2		7.1022	6.9403	5.4295	5.6837	4.9538	3.3233	3.7358	3.7911	3.6601	3.8047
3		7.8179	7.6104	7.2280	6.0766	5.6546	4.2006	4.1387	4.6700	3.8176	4.0667
4		8.8335	8.9030	8.2817	7.0014	5.9816	4.4027	5.1975	5.1425	4.1977	4.4922
5		9.8080	9.1919	8.4201	7.9682	7.2644	5.0634	5.1639	4.8035	5.5495	5.5627
6		12.3480	9.2838	9.6795	8.8177	7.6413	6.0545	5.1822	6.1263	5.7053	5.6554
7		10.7734	11.1139	9.5343	9.7030	9.3939	5.4592	5.6224	5.5001	6.7492	7.3288
8		11.2209	10.9357	10.3834	9.0716	7.9839	5.8005	6.5001	6.0242	5.6473	5.7731
9		13.3989	12.1643	9.9795	10.4110	9.1066	7.4167	6.4201	6.3731	6.4547	7.1052
10		13.0978	12.7918	11.3914	10.9843	10.3273	6.6536	7.2401	7.1739	7.3257	7.2169
11		13.3882	11.2289	12.5175	12.5353	9.8387	6.6060	6.7115	6.9667	8.0051	6.7951
12		15.1617	13.4001	12.3095	11.2742	10.5560	6.9564	6.5237	7.4666	8.0105	7.5797
13		17.1781	12.4525	11.8663	12.4799	11.0325	7.9928	7.4223	7.3339	7.5196	7.5275
14		16.0103	13.3144	12.2509	11.6300	11.6842	7.4660	8.2465	7.4890	7.6319	9.0704

Notes to Table A.1. The Critical Values are obtained by Monte Carlo simulation using 50,000 Monte Carlo replications.

Figure 1 Panel A

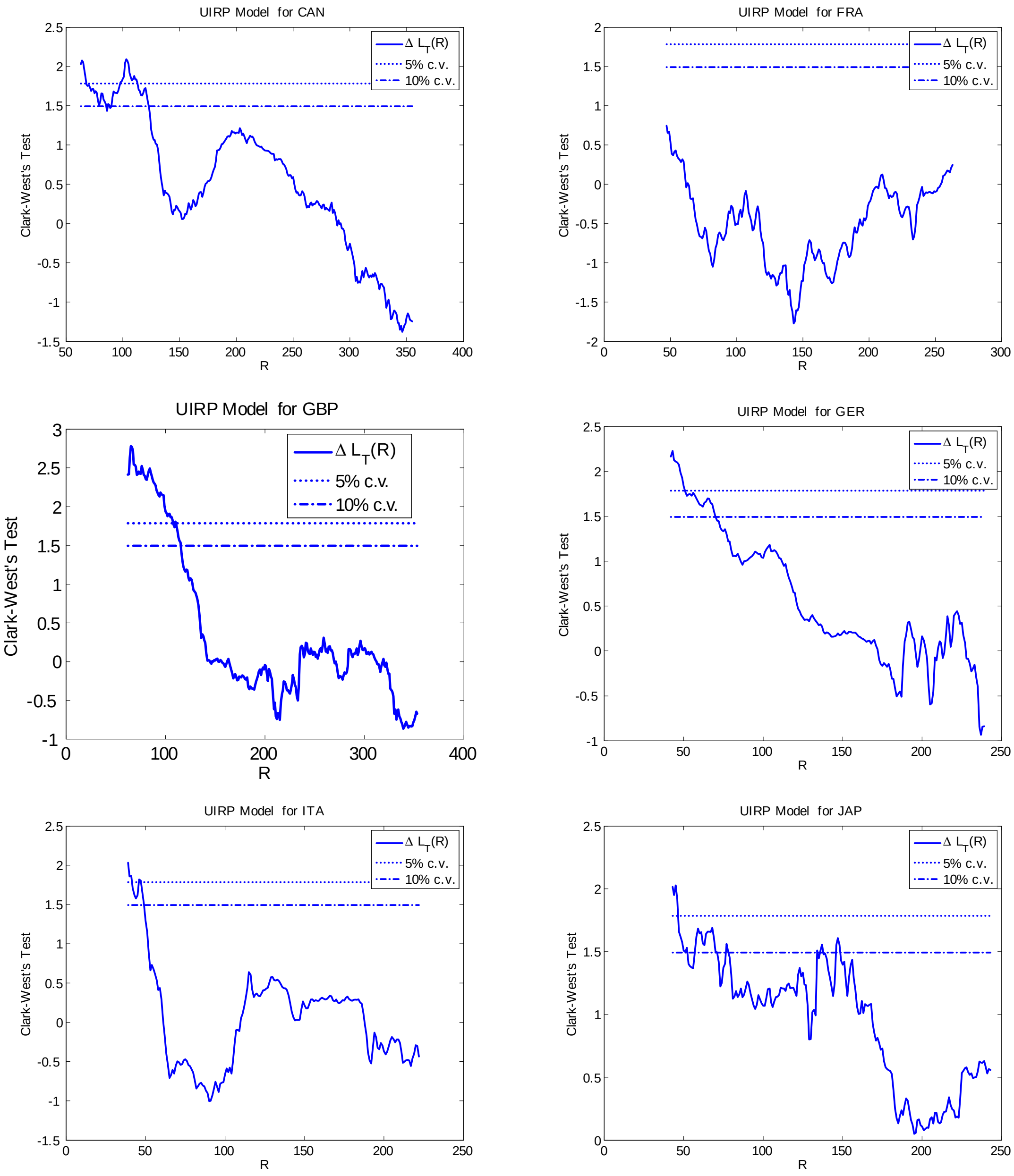


Figure 1 Panel B

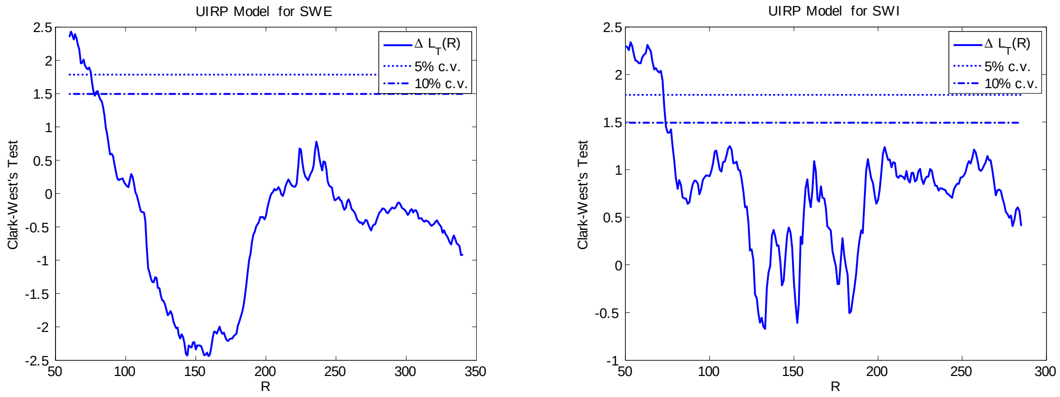
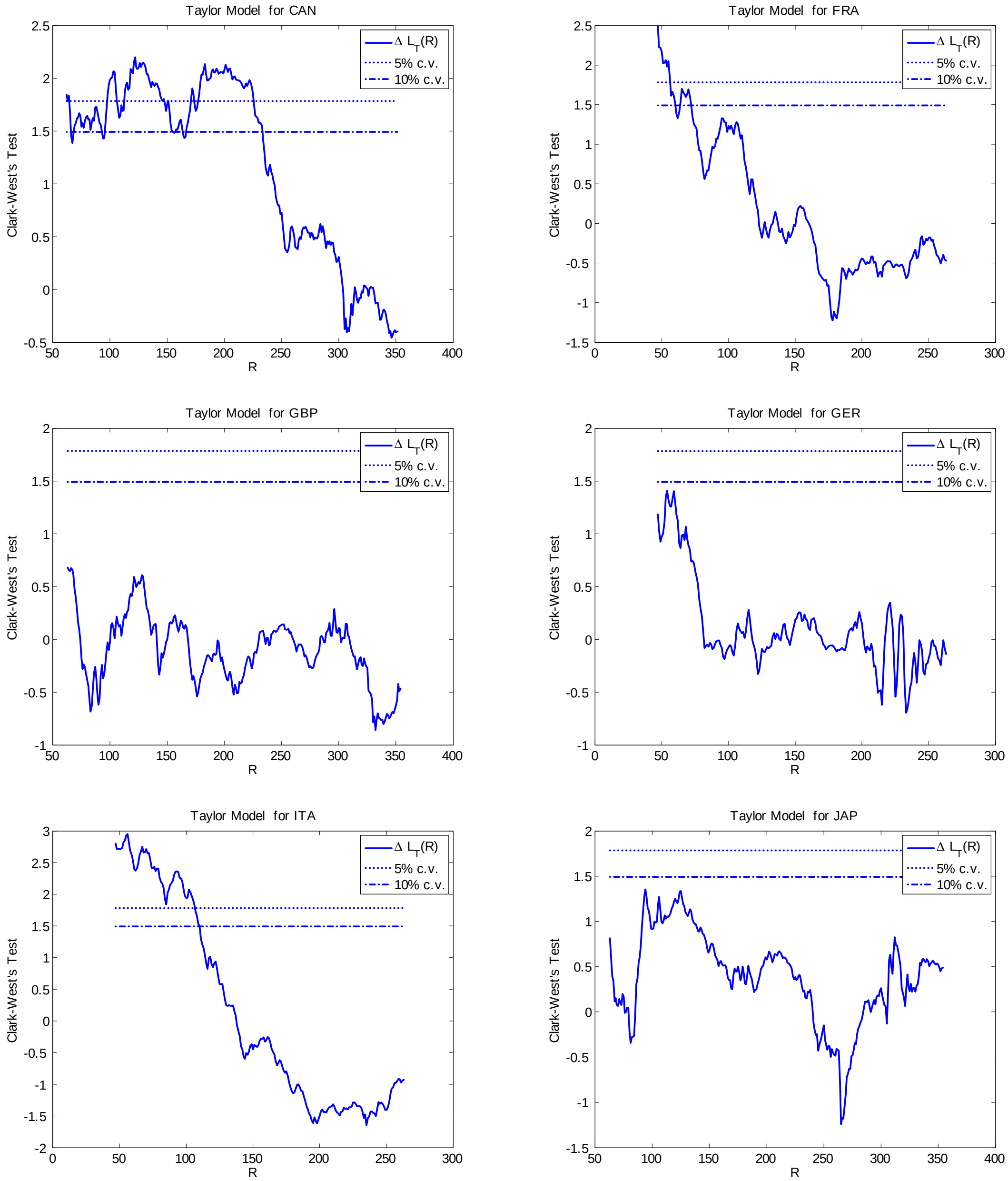


Figure 1 plots the estimated Clark and West's (2007) test statistic for the window sizes that we consider (reported on the x-axis), together with 5% and 10% critical values. Countries are: Canada (CAN), France (FRA), United Kingdom (GBP), Germany (GER), Italy (ITA), Japan (JAP), Sweden (SWE) and Switzerland (SWI).

Figure 2 Panel A



Taylor Model for CAN

Taylor Model for FRA

Taylor Model for GBP

Taylor Model for GER

Taylor Model for ITA

Taylor Model for JAP

Figure 2 Panel B

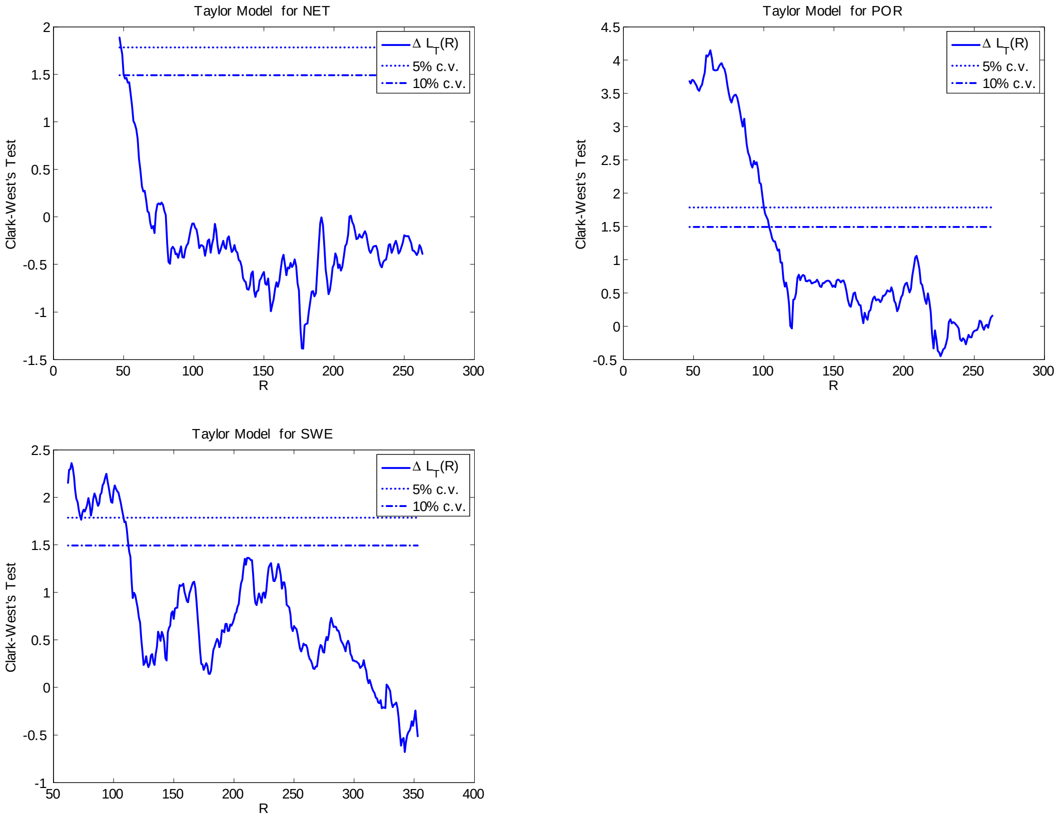


Figure 1 plots the estimated Clark and West's (2007) test statistic for the window sizes that we consider (reported on the x-axis), together with 5% and 10% critical values. Countries are: Canada (CAN), France (FRA), United Kingdom (GBP), Germany (GER), Italy (ITA), Japan (JAP), Sweden (SWE) and Switzerland (SWI), The Netherlands (NET) and Portugal (POR).