2022, Volume 45, Issue 89, 150-183 / ISSN 2304-4306

E C O N O M Í A

revistas.pucp.edu.pe/economia

www.fondoeditorial.pucp.edu.pe

# A Power Booster Factor for Out-of-Sample Tests of Predictability

Pablo Pincheira Brown[a]

[a]School of Business, Universidad Adolfo Ibáñez

✉ pablo.pincheira@uai.cl

**Abstract**

In this paper we introduce a "power booster factor" for out-of-sample tests of predictability. The relevant econometric environment is one in which the econometrician wants to compare the population Mean Squared Prediction Errors (MSPE) of two models: one big nesting model, and another smaller nested model. Although our factor can be used to improve finite sample properties of several out-of-sample tests of predictability, in this paper we focus on the widely used test developed by Clark and West (2006, 2007). Our new test multiplies the Clark and West t-statistic by a factor that should be close to one under the null hypothesis that the short nested model is the true model, but that should be greater than one under the alternative hypothesis that the big nesting model is more adequate. We use Monte Carlo simulations to explore the size and power of our approach. Our simulations reveal that the new test is well sized and powerful. In particular, it tends to be less undersized and more powerful than the test by Clark and West (2006, 2007). Although most of the gains in power are associated to size improvements, we also obtain gains in size-adjusted-power. Finally we illustrate the use of our approach when evaluating the ability that an international core inflation factor has to predict core inflation in a sample of 30 OECD economies. With our "power booster factor" more rejections of the null hypothesis are obtained, indicating a strong influence of global inflation in a selected group of these OECD countries.

**Article History:** Received: 12 February 2021 / Revised: 12 May 2021 / Accepted: 11 August 2021

**Keywords:** Time-series; Forecasting; Inference; Inflation; Exchange rates; Random walk; Out-of-sample

**JEL Classification:** C22, C52, C53, C58, E17, E27, E37, E47, F37

## Acknowledgements

I would like to thank Roberto Duncan, Enrique Martínez-García, Kenneth D. West, Esther Ruiz and two reviewers for wonderful comments. I am also grateful to participants of seminars at the Central Bank of Chile, Adolfo Ibáñez University and at the 2016 Annual Meeting of the Chilean Economics Society. Financial support from the "individual research grant 2016" of the Adolfo Ibáñez University is greatly appreciated. All errors are mine.

https://doi.org/10.18800/economia.202201.006

## 1. Introduction

In this paper we introduce a *"power-booster-factor"* for out-of-sample tests of predictability. The relevant econometric environment is one in which the econometrician wants to test for the difference between the population Mean Squared Prediction Errors (MSPE) of two models: one big nesting model, and another smaller nested model. The standard application of such comparisons is found in the exchange rate literature, where an economic model is used to generate forecasts that are compared to forecasts coming from the simple random walk.

Our *"power-booster-factor"* can be used to improve finite sample properties of several out-of-sample tests of predictability. Yet, in this paper, we focus on the widely used test developed by Clark and West (2006, 2007) (hereafter CW). We construct a new test multiplying the CW t-statistic by our *"power-booster-factor"*. The key idea relies on the fact that this factor should be close to one under the null hypothesis that the short nested model is correct, but should be greater than one under the alternative hypothesis that the big nesting model is more adequate. This new test displays two interesting features. First, standard normal critical values seem to work well, meaning that the test is correctly sized, and second, the test is relatively powerful when compared to the widely used CW test.

Out-of-sample analyses have become fairly usual in time series econometrics to compare either different forecasting methods or the adequacy of economic models. Accordingly, during the last two decades several papers have proposed different out-of-sample testing strategies. For instance, Diebold and Mariano (1995) and West (1996) are leading articles in this literature.

When the objective is to compare population MSPE of two models, and one of them is nested in the other, a vast literature has documented that the traditional methods proposed by Diebold and Mariano (1995) and West (1996) are inadequate, see for instance West (1996, 2006). In particular McCracken (2007) derives the correct asymptotic distribution of traditional comparisons of MSPE between nested models concluding that, in general, usual tests are not normal. Moreover, McCracken (2007) provides the asymptotic distribution of four widely used statistics to compare population MSPE in nested environments. The extension to direct multistep ahead forecasts is made in Clark and McCracken (2005). In general terms the tests follow a non-standard distribution.

An alternative approach is presented by CW, who show that the asymptotic distribution of a simple encompassing t-statistic is well approximated by a standard normal distribution under the null hypothesis. In the particular case in which the null model posits a martingale process for the predictand and estimates of the parameters are updated in rolling windows, CW shows that the correct asymptotic distribution is indeed standard normal.

One important shortcoming of out-of-sample analyses is the need for splitting the available sample in two shares: one for estimation and one for forecast evaluation. One undesired consequence of this approach is the reduced number of observations for parameter estimation. This problem is typically associated to low power of out-of sample tests, see for instance Inoue and Kilian (2005). In addition, in many applications it is possible to think that the relevant alternative and null models are relatively close to each other. For instance, given that in relatively

efficient asset markets we could expect little or no predictability of returns, it is critical to rely on high power tests, so they can detect this presumably little evidence against the null hypothesis.

The joint use of the CW test and our *"power-booster-factor"* allows us to propose a new test with relatively high power based on asymptotically normal critical values, which are very simple to use.

We use Monte Carlo simulations to explore size, power and size-adjusted-power of our new test when forecasting one-step-ahead. We are interested in its performance both absolutely and relative to CW, which is the other usual asymptotically normal test used in nested environments. In our simulations, we calibrate our parameters and sample sizes to macro applications based on monthly exchange rates or monthly CPI inflation.

Simulation results reveal that our new test behaves as expected: it is, in general, correctly sized and more powerful than CW. Notice, however, that improvements in size-adjusted-power are moderate, and gains in power are mostly induced by our new test being less undersized than CW.

While our test may display adequate size and high power, there are plenty of subtleties that deserve mentioning: First, our approach tends to be slightly undersized when carrying out inference at the 10% level, but it is a little oversized at the 1% level. Second, in applied work the researcher needs to make a decision about one free parameter. We provide some suggestions on how to pick that parameter, but more should be done in the future.

We emphasize that we are testing equal population forecasting ability. In other words we use forecast comparisons as a model evaluation technique. We leave as an extension for future research the connection of our test with procedures to obtain good forecasts on a given sample. See Giacomini and White (2006) for an interesting discussion about the differences in evaluating forecasting methods and models. Finally we illustrate the use of our approach when evaluating the ability that an international core inflation factor has to predict core inflation in a sample of 30 OECD economies. With our *"power-booster-factor"* more rejections of the null hypothesis are obtained, indicating a strong influence of global inflation in a selected group of these OECD countries.

The rest of the paper is organized as follows. Section 2 outlines the general econometric environment, the CW test, the "*power-booster-factor*" and the construction of our new test. Section 3 shows some asymptotic and finite sample results and observations. In Section 4 we describe our DGPs and the simulation setup. Results of the Monte Carlo experiment are shown in Section 5. Section 6 illustrates the use of our test in an empirical application and Section 7 concludes.

## 2. Econometric Setup and Forecast Evaluation Framework

### 2.1 Basic Econometric Setup

We use a linear econometric setup considering nested specifications for a scalar dependent variable $y_{t+1}$ as follows:

$$\text{(model 1: null model)}: \qquad y_{t+1} = X_t'\beta + e_{1t+1}, \tag{2.1}$$

$$\text{(model 2: alternative model)}: \qquad y_{t+1} = X_t'\beta + Z_t'\gamma + e_{2t+1}, \tag{2.2}$$

where $e_{1t+1}$ and $e_{2t+1}$ are zero mean martingale difference sequences meaning that $E(e_{it+1}|F_t) = 0$ for $i = 1, 2$. Here $F_t$ represents the sigma-field generated by current and past values of $X_t$, $Z_t$ and $e_{it}$ for $i = 1, 2$.

We are interested in evaluating the following null hypothesis: $H_0 : \gamma = 0$. When this hypothesis is true, model 2 and model 1 are the same. This means that in population, forecasts, forecast errors and Mean Squared Prediction Errors (MSPE) are the same in both models. Under the alternative, $\gamma \neq 0$, and forecasts will be different in both models. In particular, since model 2 includes relevant information for explaining $y_t$, the population forecasts from model 2 will be superior to those of model 1, meaning that model 2 will have a lower MSPE than model 1.

We focus on the evaluation of our proposed test when comparing one-step-ahead forecasts. Let $\hat{y}_{1,t+h|t}$ and $\hat{y}_{2,t+h|t}$ represent $h$ period ahead forecasts from each of the two models. Let $\hat{\beta}_{1t}$ be a least squares estimate of model 1 that only uses data up to period $t$, with $\hat{\beta}_{2t}$ and $\hat{\gamma}_{2t}$ the model 2 counterparts. Then one-step-ahead forecasts and forecasts errors are given by the following expressions

$$\hat{y}_{1,t+1|t} = X_t'\hat{\beta}_{1t}, \qquad \hat{y}_{2,t+1|t} = X_t'\hat{\beta}_{2t} + Z_t'\hat{\gamma}_{2t}. \tag{2.3}$$

$$\hat{e}_{1,t+1|t} \equiv y_{t+1} - X_t'\hat{\beta}_{1t}, \qquad \hat{e}_{2,t+h|t} \equiv y_{t+1} - X_t'\hat{\beta}_{2t} - Z_t'\hat{\gamma}_{2t}. \tag{2.4}$$

## 2.2 The Test by Clark and West

As our *"power-booster-factor"* is heavily based on the CW test, it will be useful to explain in some detail the rationale behind this widely used test. To that end we need to describe our out-of-sample exercises. Let us assume that we have a total of $T + 1$ observations on $y_t$. The end point of the first sample used to estimate regression parameters is observation $R$. We generate a sequence of $P$ one-step-ahead forecasts estimating the models in either rolling windows of fixed size $R$ or recursive windows of size equal or greater than $R$.

For rolling windows, to generate the first set of forecasts we estimate our models with the first $R$ observations of our sample. Thus, these forecasts are built with information available only at time $R$ and are compared to the observation $y_{R+1}$. Next, we estimate our models with the second rolling window of size $R$ that includes observations 2 through $R + 1$. These forecasts are compared to observation $y_{R+2}$. We continue until the last forecasts are built using the last $R$ available observations for estimation. These forecasts are compared to observation $y_{T+1}$.

When recursive or expanding windows are used instead, the only difference with the procedure described in the previous paragraph relates to the size of the estimation windows. In the recursive scheme, the estimation window size grows with the number of available observations for estimation. For instance, the first forecast is constructed estimating the models in a window of size $R$, whereas the final forecast is constructed based on models estimated in a window of size $T$. Thus, we generate a total of $P$ forecasts, with $P$ satisfying $R + (P - 1) + 1 = T + 1$. So $P = T + 1 - R$.

Sample estimates of Mean Squared Prediction Errors (MSPE) from the two models are

$$\hat{\sigma}_1^2 = \frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2 \tag{2.5}$$

$$\hat{\sigma}_2^2 = \frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{2,t+1|t}^2. \tag{2.6}$$

Under the null, the population MSPE of both models is the same: $\sigma_1^2 = \sigma_2^2$; under the alternative, the population MSPE of the bigger model should be lower than the population MSPE of the smaller model: $\sigma_1^2 > \sigma_2^2$. Specifically, construction of CW starts by producing an adjusted estimate of the MSPE from model 2,[1]

$$\hat{\sigma}_2^2 - adj. = \frac{1}{P}\sum_{t=R}^{R+P-1} \left[\hat{e}_{2,t+1|t}^2 - \left(\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t}\right)^2\right]. \tag{2.7}$$

Now define $\hat{V}$ to be a consistent estimate of the long run variance of $\hat{e}_{1,t+1|t}^2 - [\hat{e}_{2,t+1|t}^2 - (\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t})^2]$. The CW test relies on the following t-statistic

$$\frac{\hat{\sigma}_1^2 - \left(\hat{\sigma}_2^2 - adj.\right)}{\sqrt{\hat{V}}}. \tag{2.8}$$

Notice that

$$\hat{e}_{2,t+1|t}^2 = \left(y_{t+1} - \hat{y}_{2,t+1|t}\right)^2 = \left((y_{t+1} - \hat{y}_{1,t+1|t}) + (\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t})\right)^2. \tag{2.9}$$

Therefore

$$\left[\hat{e}_{2,t+1|t}^2 - (\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t})^2\right] = \left[\hat{e}_{1,t+1|t}^2 + 2\hat{e}_{1,t+1|t}(\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t})\right]. \tag{2.10}$$

Or, equivalently,

$$\left[\hat{e}_{2,t+1|t}^2 - (\hat{y}_{1,t+1|t} - \hat{y}_{2,t+1|t})^2\right] = \left[\hat{e}_{1,t+1|t}^2 - 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t})\right]. \tag{2.11}$$

Consequently, (2.7) could also be written as follows

$$\hat{\sigma}_2^2 - adj. = \frac{1}{P}\sum_{t=R}^{R+P-1} \left[\hat{e}_{1,t+1|t}^2 - 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t})\right]. \tag{2.12}$$

From (2.12) it is straightforward to see that the numerator of the CW t-statistic is equal to

$$\hat{\sigma}_1^2 - \left(\hat{\sigma}_2^2 - adj.\right) = \frac{2}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}\left(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t}\right). \tag{2.13}$$

[1] Clark and West (2006, 2007) explain the logic leading to this adjustment.

## 2.3 The Power-Booster-Factor

Let us consider the following Percentage Difference in Mean Squared Prediction Errors (PDMSPE):

$$\text{PDMSPE} = \frac{\sigma_1^2 \; - \; \sigma_2^2}{\sigma_1^2}.$$

At the population level this percentage difference (relative to model 1) is expected to be zero when the null is true and positive otherwise. At the sample level, however, this is not always the case given that the extra parameters in model 2 inflate the MSPE of that model with additional estimation error. Following the logic behind Clark and West (2006, 2007) we should adjust $\hat{\sigma}_2^2$ accordingly. This means to consider the following Sample PDMSPE ratio:

$$\widehat{\text{PDMSPE}} = \frac{\hat{\sigma}_1^2 - [\hat{\sigma}_2^2 - adj.]}{\hat{\sigma}_1^2}.$$

Using (2.5) and (2.12) we see that

$$\widehat{\text{PDMSPE}} = \frac{\frac{2}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}\left(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t}\right)}{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2}.$$

Given that this term represents a percentage variation we simply need to add 1 to get our *"power-booster-factor"*:

$$1 + \widehat{\text{PDMSPE}} = 1 + \frac{\frac{2}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}\left(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t}\right)}{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2}.$$

Which is equivalent to:

$$\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2} = \frac{\frac{1}{P}\sum_{t=R}^{R+P-1}\left[\hat{e}_{1,t+1|t}^2 + 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t})\right]}{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2}. \tag{2.14}$$

Notice that the numerator in (2.14) corresponds to $\hat{\sigma}_1^2$ plus the CW core statistic. This is important, because this last statistic has a different behavior under the null and alternative hypotheses. When the null hypothesis is true, we expect the core CW statistic to be close to zero, therefore, under the null

$$\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj)}{\hat{\sigma}_1^2} \approx \frac{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2}{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2} = 1. \tag{2.15}$$

Under the alternative, the core CW statistic should be positive, which implies

$$\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2} = \frac{\frac{1}{P}\sum_{t=R}^{R+P-1}\left[\hat{e}_{1,t+1|t}^2 + 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t})\right]}{\frac{1}{P}\sum_{t=R}^{R+P-1} \hat{e}_{1,t+1|t}^2} > 1. \tag{2.16}$$

In the particular case in which the null model is a simple martingale in difference process and parameter estimates are updated in rolling windows, expression (2.14) will converge in probability to 1 when the null hypothesis is true. We will see this with formal arguments in Section 3.

It is important to notice, that we could raise expression (2.14) to some positive scalar $\lambda$ as follows

$$\left(\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2}\right)^{\lambda} = \left(\frac{\frac{1}{P}\sum_{t=R}^{R+P-1}[\hat{e}_{1,t+1|t}^2 + 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+1|t} - \hat{e}_{2,t+1|t})]}{\frac{1}{P}\sum_{t=R}^{R+P-1}\hat{e}_{1,t+1|t}^2}\right)^{\lambda}. \tag{2.17}$$

This simple transformation preserves the basic properties under the null and alternative hypotheses. Expression (2.17) introduces our "*power-booster-factor*". It depends on the parameter $\lambda$, which should play no role aymptotically under the null hypothesis. We will explore via simulations the different behavior of our *"power-booster-factor"* as a function of $\lambda$ in Section 5.

## 2.4 Our New Test

We propose to multiply the CW t-statistic by the factor in (2.17) to construct an asymptotically normal test. As we will see in the next section, in the particular case in which the null model is a simple martingale in difference process and parameter estimates are updated in rolling windows, expression (2.17) will converge in probability to 1 when the null hypothesis is true. This, Slutsky's theorem plus asymptotic normality of the test by Clark and West (2006) ensures asymptotic normality for our approach. In the case of the test in Clark and West (2007) which is not normal, we rely on the good behavior of the normal approximation described by simulations in that paper, and many others, to use normal critical values for our test as well.

Notice that under the null hypothesis both tests, ours and CW, should be asymptotically the same, but under the alternative hypothesis, the factor in (2.17) should be greater than 1, and therefore it should boost and improve the power of the CW test. Furthermore, the higher the CW core statistic is, the higher the factor in (2.17) is, which suggest that gains in power should be greater at the 5% level than at the 10% significance level, and also should be greater at the 1% than at the 5% significance level.

Notice also that our procedure can be easily extended to general h-steps-ahead forecasts with the same logic as well. Using the same models 1 and 2, let us consider $\hat{e}_{1,t+\mathrm{h}|t} \equiv \hat{y}_{1,t+\mathrm{h}|t} - \hat{X}'_{t+\mathrm{h}-1|t}\hat{\beta}_{1t}$ and $\hat{e}_{2,t+\mathrm{h}|t} \equiv \hat{y}_{2,t+\mathrm{h}|t} - \hat{X}'_{t+\mathrm{h}-1|t}\hat{\beta}_{2t} - \hat{Z}'_{t+\mathrm{h}-1|t}\hat{\gamma}_{2t}$ the forecast errors of both models at horizon h.

In this context, our power booster factor is defined as follows

$$\mathrm{PBF(h)} = \left(\frac{\frac{1}{\mathrm{P(h)}}\sum_{t=R}^{R+\mathrm{P(h)}-1}[\hat{e}_{1,t+h|t}^2 + 2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+h|t} - \hat{e}_{2,t+h|t})]}{\frac{1}{\mathrm{P(h)}}\sum_{t=R}^{R+\mathrm{P(h)}-1}\hat{e}_{1,t+h|t}^2}\right)^{\lambda}.$$

Where P(h) denotes the number of h-step ahead under consideration. Under the null hypothesis the term

$$\frac{1}{\mathrm{P(h)}}\sum_{t=R}^{R+\mathrm{P(h)}-1}\left[2\hat{e}_{1,t+1|t}(\hat{e}_{1,t+h|t} - \hat{e}_{2,t+h|t})\right] \tag{2.18}$$

is expected to be close to zero. Therefore PBF(h) should be close to one. Under the alternative hypothesis we should expect the term (2.18) to be positive, which implies a value greater than one for PBH(h). For one-sided-tests this means higher t-statistics and hence, more power.[2]

[2] A critical aspect to take into consideration is that the *"power-booster-factor"* induces some changes in size as

## 3. Asymptotic and Finite Sample Behavior of our Approach

### 3.1 Simple Asymptotic Theory

Here we provide a formal asymptotic analysis for our new test in the particular case in which the null model is a simple martingale in difference process and parameter estimates are updated in rolling windows, as in Clark and West (2006). This means that in (2.1) and (2.2) we are considering the special case $\beta = 0$. So the models are

$$y_{t+1} = e_{1t+1}, \tag{3.1}$$

$$y_{t+1} = Z_t'\gamma + e_{2t+1}, \tag{3.2}$$

Under the null, $\gamma = 0$, so in population the subscripts 1 and 2 are no longer necessary. Therefore we could write

$$y_{t+1} = e_{t+1}, \tag{3.3}$$

In (3.2), let $\hat{\gamma}_t$ denote an estimate of $\gamma$ that relies on data going from $t - R + 1$ to $t$. We have

$$\hat{y}_{1,t+1|t} = 0, \hat{e}_{1,t+1|t} = y_{t+1}, \hat{e}_{2,t+1|t} = y_{t+1} - Z_t'\hat{\gamma}_t, \tag{3.4}$$

$$\hat{e}^2_{1,t+1|t} - \hat{e}^2_{2,t+1|t} = y^2_{t+1} - \left(y_{t+1} - Z_t'\hat{\gamma}_t\right)^2 = 2y_{t+1}Z_t'\hat{\gamma}_t - \left(Z_t'\hat{\gamma}_t\right)^2. \tag{3.5}$$

Thus the numerator of the CW statistic is

$$\hat{\sigma}_1^2 - \left(\hat{\sigma}_2^2 - adj.\right) = \frac{2}{P}\sum_{t=R}^{R+P-1} y_{t+1}Z_t'\hat{\gamma}_t. \tag{3.6}$$

And our "power-booster-factor" factor from (2.17) looks as follows:

$$\left(\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2}\right)^{\lambda} = \left(\frac{\frac{1}{P}\sum_{t=R}^{R+P-1}[y^2_{t+1} + 2y_{t+1}Z_t'\hat{\gamma}_t]}{\frac{1}{P}\sum_{t=R}^{R+P-1} y^2_{t+1}}\right)^{\lambda}. \tag{3.7}$$

Under the null hypothesis, since $y_{t+1} = e_{t+1}$ is a martingale difference sequence, and $\hat{\gamma}_t, Z_t'\hat{\gamma}_t$ relies only on data that ends in $t$, $E(y_{t+1}Z_t'\hat{\gamma}_t) = E(e_{t+1}Z_t'\hat{\gamma}_t) = 0$. Thus the expectation of the numerator of the CW statistic is

$$E(\hat{\sigma}_1^2 - \hat{\sigma}_2^2 - adj.) = 0. \tag{3.8}$$

Given that $e_{t+1}Z_t'\hat{\gamma}_t$ is also a martingale difference sequence we could use a standard central limit theorem for martingale processes to show asymptotic normality for the CW statistic. We need some additional assumptions to show that the *"power-booster-factor"* converges in probability to 1 under the null hypothesis.

Let us consider the following assumptions:

$$E(e_{t+1}Z_t'\hat{\gamma}_t)^2 > 0 \text{ and } \lim_{P\to\infty}\frac{1}{P}\sum_{t=R}^{R+P-1} E(e_{t+1}Z_t'\hat{\gamma}_t)^2 V^* > 0; \tag{3.9}$$

well. Therefore, our approach is recommendable when used with slightly undersized tests. Simulations completed by Pincheira and West (2016) show that Clark and West is indeed undersized at short horizons of h=1, h=2 and h=3 steps ahead. Consequently, our approach is expected to be adequate at these horizons as well.

$$E|e_{t+1}Z_t'\hat{\gamma}_t|^{2d} < M < +\infty \text{ for } d > 1 \text{ for all } t; \tag{3.10}$$

$$\frac{1}{P}\sum_{t=R}^{R+P-1}(e_{t+1}Z_t'\hat{\gamma}_t)^2 \xrightarrow{Pr} V^* \text{ as } P \text{ goes to infinity;} \tag{3.11}$$

$$\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2 \xrightarrow{Pr} \tau^* > 0 \text{ as } P \text{ goes to infinity;} \tag{3.12a}$$

$$\left(\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2\right)^{-1} \text{ is bounded in probability.} \tag{3.12b}$$

Assumptions (3.9), (3.10) and (3.11) are required for the central limit to hold true. See Hamilton (1994) for details. Therefore we have that

$$\frac{\sqrt{P}}{P}\sum_{t=R}^{R+P-1}(e_{t+1}Z_t'\hat{\gamma}_t) \to N(0, V^*). \tag{3.13}$$

Similarly, assumption (3.10) implies that the law of large numbers holds for $(e_{t+1}Z_t'\hat{\gamma}_t)$. Meaning that

$$\frac{1}{P}\sum_{t=R}^{R+P-1}(e_{t+1}Z_t'\hat{\gamma}_t) \to 0. \tag{3.14}$$

This convergence is achieved almost surely, and therefore it is also satisfied in probability. See White (2001) for further details.

Assumptions (3.12a) and (3.12b) are different alternatives required for our *"power-booster-factor"* to converge in probability to 1 under the null hypothesis. Assumption (3.12a) is more restrictive than (3.12b) because the sequence of the sample average of $e_{t+1}^2$ is required to converge in probability. Assumption (3.12b) does not require convergence. It requires the sample average of $e_{t+1}^2$ to be far away from zero, which is a reasonable requirement, given that this is the sample average of a sequence of positive random variables.

Using assumption (3.12a) our *"power-booster-factor"* converges in probability to 1 under the null hypothesis, given that expression (3.14) holds true. To complete the argument we advocate the continuous mapping theorem applied to the power function $f(x) = x^\lambda$. Therefore:

$$\left(\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2}\right)^\lambda = \left(\frac{\frac{1}{P}\sum_{t=R}^{R+P-1}[e_{t+1}^2 + 2e_{t+1}Z_t'\hat{\gamma}_t]}{\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2}\right)^\lambda \xrightarrow{Pr} \left(\frac{\tau^*}{\tau^*}\right)^\lambda = 1. \tag{3.15}$$

We arrive at the same conclusion using assumption (3.11b) but writing (3.7) in a slightly different way:

$$\left(\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2}\right)^\lambda = \left(\frac{\frac{1}{P}\sum_{t=R}^{R+P-1}[e_{t+1}^2 + 2e_{t+1}Z_t'\hat{\gamma}_t]}{\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2}\right)^\lambda = \left(1 + \frac{\frac{1}{P}\sum_{t=R}^{R+P-1}[2e_{t+1}Z_t'\hat{\gamma}_t]}{\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2}\right)^\lambda. \tag{3.16}$$

The joint use of (3.14) and assumption (3.12b) implies that (3.16) converges in probability to 1 under the null hypothesis.[3]

[3]Here we are using the following result: If $Y_n$ converges in probability to zero, and $X_n$ is bounded in probability, then the product $Y_nX_n$ converges in probability to zero as well.

Our proposal is to multiply the CW t-statistic by our *"power-booster-factor"*. Asymptotic normality under the null hypothesis follows from asymptotic normality of CW, the *"power-booster-factor"* converging to one in probability, plus the application of Slutsky's theorem.

As usual in this literature, assumption (3.9) rules out the use of recursive or expanding windows in the out-of-sample analysis, so for asymptotic normality to hold true, we rely on rolling regressions.

As mentioned before, under the alternative hypothesis we expect the CW core statistic to be positive. This means a *"power-booster-factor"* greater than one.

$$\left(\frac{2\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.)}{\hat{\sigma}_1^2}\right)^{\lambda} = \left(1 + \frac{\frac{2}{P}\sum_{t=R}^{R+P-1}[2e_{t+1}Z_t'\hat{\gamma}_t]}{\frac{1}{P}\sum_{t=R}^{R+P-1} e_{t+1}^2}\right)^{\lambda} > 1. \tag{3.17}$$

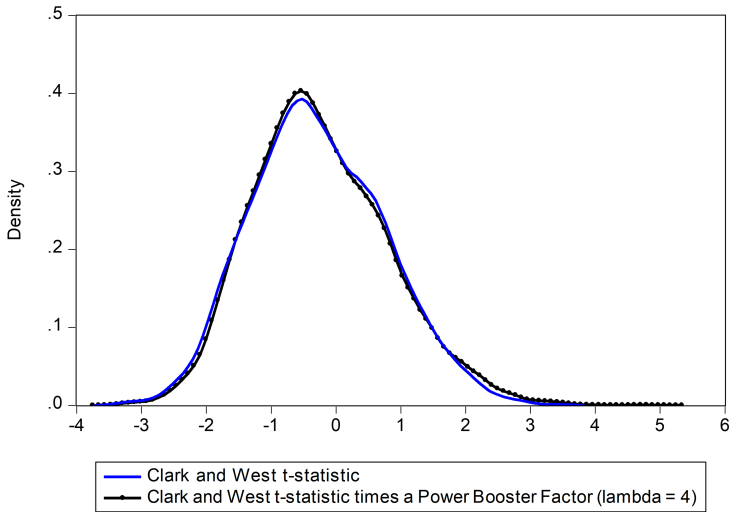
For one sided test, the implication is that our approach should have more power.

A final point is worth mentioning. When one uses the recursive scheme or when $\beta \neq 0$ in (2.2), so that the null model includes at least one regressor, we do not have a proof of asymptotic normality with or without our *"power-booster-factor"*. As a matter of fact, the asymptotic distribution of the CW statistic is not normal. Clark and McCracken (2001) derive the correct asymptotic distribution of the CW test when one-step-ahead forecasts are used, and Clark and McCracken (2005) do the same when longer horizon forecasts are constructed via the direct method. In the first paper it is shown that the resulting asymptotic distribution of the CW test in general is not standard. In fact it is a functional of Brownian motions depending on the number of excess parameters of the nesting model, the limit of the ratio P(h)/$R$ and the scheme used to update the estimates of the parameters in the out-of-sample exercise (rolling, recursive or fixed). In the second paper, Clark and McCracken (2005) provide a generalization of their results for multistep ahead forecasts. Unfortunately, the resulting asymptotic distribution of the CW statistic is again a functional of Brownian motions but now depending on nuisance parameters.

Differing from the previous work of Clark and McCracken (2001, 2005), one of the key contributions of CW is to show via simulations that normal critical values are indeed adequate in a variety of settings. They show that the cost of approximating the correct critical values by standard normal ones is in general low: it produces a little undersized test. Furthermore, simulations completed by Clark and McCracken (2013) and Pincheira and West (2016) are consistent with the view that the CW statistic can reasonably be thought of as approximately normal. We will see via simulations in the following sections that our approach also seems to work well with standard normal critical values in a variety of settings.

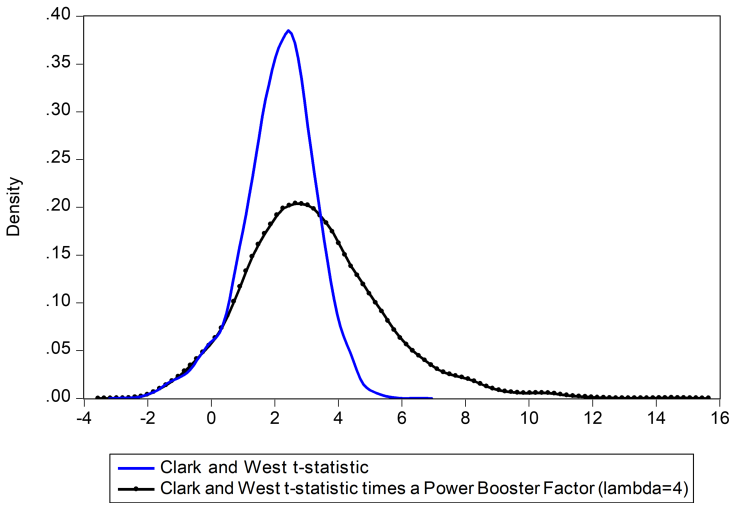
## 3.2 Finite Sample Behavior

In finite samples of size typically available in macroeconomics, it is not always the case that the asymptotic theory will be useful to explain the behavior of our test. In particular, when the number of forecasts $P$ is moderate or small, the multiplication of our *"power-booster-factor"* and the CW t-statistic has a skewed distribution: it tends to have a heavier right tail. See Figures 1-2 below. This is due to the fact that our test is a nonlinear function of the CW core statistic: When this core statistic is close to zero, our test is close to CW, but when the CW core statistic




*Notes*: Data for Figures 1 and 2 come from 5000 replications of DGP 3, using the recursive scheme with parameters $P = 120$ and $\lambda = 4$. See Section 4 next for a description of our Monte Carlo simulations.

**Figure 1.** Kernel densities of the Clark and West t-statistic and our new test distributions under the null hypothesis.




*Notes*: See notes to Figure 1.

**Figure 2.** Kernel densities of the Clark and West t-statistic and our new test distributions under the alternative hypothesis.

is large; our test would be even larger. This feature has implications in terms of size and power. In terms of power, the implication is that we expect our test to show more power relative to CW at the 1% significance level than at the 5% significance level. Similarly, our test should show more power relative to CW at the 5% level than at the 10% level. In terms of size, the same nonlinear dynamics holds true, so our test should display higher size than CW. This seems not to be a serious problem given that simulations completed in Clark and West (2006, 2007) show that CW is a little undersized in finite samples. An increment in size could be beneficial if this

increment is just moderate.

A final point reflects the observation that the *"power-booster-factor"* is an increasing function of the parameter $\lambda$. This will also have implications in finite samples, meaning that the power of the test should increase with $\lambda$.[4] In terms of size, the implication is that the empirical size of our test increases with $\lambda$ as well. The recommendation here is to pick low levels for $\lambda$. We will go back to this issue in Section 5.

## 4. Monte Carlo Simulations

Our three DGPs are stimulated by empirical work in asset pricing and macroeconomics. Most driving shocks are i.i.d. normal, but in DGP 1 we also experiment with shocks displaying fat tails. In all simulations we consider both rolling and recursive samples, several values for the parameter $\lambda$ in (3.17), a single value of the initial regression sample size $R$ and four values of the number of one-step-ahead predictions $P$.

### 4.1 Experimental Design

DGP 1: Here we focus on the case where the null is a martingale model. DGP 1 is fairly similar to the first DGP in Pincheira and West (2016) and to those used in Clark and West (2006), Mankiw and Shapiro (1986), Nelson and Kim (1993), Stambaugh (1999), Campbell (2001), Tauchen (2001) and Pincheira (2013). This DGP is designed to match exchange rate series for which the martingale difference is a plausible null hypothesis and a model based on uncovered interest parity is a plausible alternative. The general setup is the following:

Null model:

$$\text{(model 1)}: \qquad y_{t+1} = e_{t+1}. \tag{4.1}$$

Alternative model:

$$\text{(model 2)}: \qquad y_{t+1} = \alpha_y + \gamma r_t + e_{t+1} \tag{4.2a}$$

$$r_{t+1} = \alpha_r + \varphi_1 r_t + \varphi_2 r_{t-1} + \ldots + \varphi_p r_{t-(p-1)} + v_{t+1}. \tag{4.2b}$$

Here, both shocks, $e_{t+1}$ and $v_{t+1}$ are independent white noise processes. While $v_{t+1}$ is assumed to be Gaussian, $e_{t+1}$ is assumed to have a $t(7)$ distribution displaying fat tails, which is a traditional feature of exchange rate returns. This simple setup maps into the notation of (2.1)-(2.2) in the following way: the term $X_t'\beta$ is set to zero and $Z_t = (1\, r_t)'$. In all our simulations, $\alpha_y = \alpha_r = \varphi_3 = \ldots = \varphi_p = 0$, so the process for $r_{t+1}$ is a driftless AR(2) model. Let

$$var(e_{t+1}) = \sigma_e^2; var(v_{t+1}) = \sigma_v^2; corr(e_{t+1}, v_{t+1}) = \rho. \tag{4.3}$$

We parameterize this as follows:

| | $\varphi_1$ | $\varphi_2$ | $\sigma_e^2$ | $\sigma_v^2$ | $\rho$ | $\gamma$, under $H_0$ | $\gamma$, under $H_A$ |
|---|---|---|---|---|---|---|---|
| DGP1 | 1.19 | -0.25 | $(1.75)^2$ | $(0.075)^2$ | 0 | 0 | -1 |

(4.4)

[4] In Appendix B, we use the delta method to show how our *"power-booster-factor"* induces higher power in a simple t-statistic testing the zero mean null hypothesis in a sample of i.i.d. observations.

In DGP 1, the null forecast (model 1) imposes $\alpha_y = \gamma = 0$, thus assuming $y_{t+1} = e_{t+1}$. The null yields simply the martingale difference or "no change" forecast of 0 for all $t$ and all forecasting horizons. (In terms of the notation above, $\hat{y}_{1,t+1|t} = 0$ for all $t$.) In DGP 1, the alternative forecast (model 2) is obtained from equation (4.2a), i.e. from a regression of $y_{t+1}$ on the first lag of $r_t$ and a constant. For the alternative, we compute forecasts using OLS estimates of our parameters, so they have the following shape

$$\hat{y}_{t+1|t} = \hat{\alpha}_{yt} + \hat{\gamma}_t r_t. \tag{4.5}$$

Here, the $t$ subscripts on the coefficients $\hat{\alpha}_{yt}$ and $\hat{\gamma}_t$ emphasize that they are estimated from a sample that ends at date $t$.

The parameterization in DGP 1 is based on estimates from the exchange rate application considered in the empirical work reported in Clark and West (2006), in which $y_{t+1}$ is the monthly percentage change in a US dollar bilateral exchange rate and $r_t$ is the corresponding interest differential. The parameters are obtained from monthly data. For this DGP we use an initial estimation window of 120 observations ($R = 120$) and report results for several different number of predictions: $P = 120$, 240, 360 and 1000. The initial window of $R = 120$ corresponds to a sample size of 10 years, the values $P = 120$, 240 and 360 represents 10, 20 and 30 years of predictions. We also consider the case in which $P = 1000$ to analyze the asymptotic behavior of our approach.

DGP 2: Our second DGP corresponds to the very same DGP 3 in Pincheira and West (2016). This DGP is motivated by the literature on commodity currencies. Our DGP 2 is designed to match monthly returns of the Non-Fuel Commodity Price Index of the IMF, $y_{t+1}$, and monthly returns of three commodity currencies versus the U.S. dollar: $r_{1t}$ = Australia, $r_{2t}$ = South Africa and $r_{3t}$ = Chile. According to Chen et al. (2010) commodity currencies should have the ability to predict commodity returns. The null model is as follows:

$$\text{(model 1)}: \qquad y_{t+1} = \alpha_y + \delta y_t + e_{t+1}. \tag{4.6}$$

The alternative model looks as follows:

$$\text{(model 2)}: \qquad y_{t+1} = \alpha_y + \gamma_1 r_{1t} + \gamma_2 r_{2t} + \gamma_3 r_{3t} + \delta y_t + e_{t+1} \tag{4.7a}$$

$$r_{it+1} = \alpha_{ir} + \varphi_i r_{it} + v_{it+1},\ i = 1, 2, 3. \tag{4.7b}$$

In the notation of (2.1)-(2.2), $X_t = y_t$ and $Z_t = (1\ r_{1t}\ r_{2t}\ r_{3t})'$. We consider the following parameters:

$$\alpha_y = \alpha_{1r} = \alpha_{2r} = \alpha_{3r} = 0,\ \delta = 0.3,\ \varphi_1 = \varphi_2 = 0.33,\ \varphi_3 = 0.5; \tag{4.8}$$
$$\text{under } H_0,\ \gamma_1 = \gamma_2 = \gamma_3 = 0;\ \text{under } H_A,\ \gamma_1 = -0.06;\ \gamma_2 = -0.015;\ \gamma_3 = -0.06.$$

These parameters are designed to match 1990–2015 monthly data, with the three currencies monthly average of daily values. The variance-covariance structure of the shocks ($e_{t+1}$, $v_{1t+1}$, $v_{2t+1}$, $v_{3t+1}$) is given by $10^{-3}$ times the following matrix

$$\begin{pmatrix} 0.536 & -0.296 & -0.229 & -0.221 \\ -0.296 & 0.666 & 0.352 & 0.251 \\ -0.229 & 0.352 & 1.09 & 0.251 \\ -0.221 & 0.251 & 0.251 & 0.478 \end{pmatrix}$$

We consider an initial estimation window of 120 observations ($R = 120$) and several different number of predictions: $P = 85$, 170, 340 and 1000.

DGP 3: Our last DGP is based on recent work exploring the predictive linkages between domestic and international inflation. Several articles analyze this relationship concluding that the linkage is important, both at the core and headline level, at least for some countries. See for instance Ciccarelli and Mojon (2010), Duncan and Martínez-García (2015), Kabukcuoglu and Martínez-García (2018), Morales-Arias and Moura (2013), Hakkio (2009), Pincheira and Gatty (2016) and Medel et al. (2016). For clarity, we relabel $y_t$ as $\pi_t^{core}$ and $r_t$ as $\pi_t^{CIIF}$, where CIIF stands for Core International Inflation Factor. The DGP is as follows. Let $e_t$ and $v_t$ be i.i.d. shocks.

$$\text{(model 1)}: \qquad \pi_{t+1}^{core} = \alpha_\pi + \varphi_\pi \pi_t^{core} + e_{t+1} \tag{4.9}$$

$$\text{(model 2)}: \qquad \pi_{t+1}^{core} = \alpha_\pi + \varphi_\pi \pi_t^{core} + \gamma_1 \pi_t^{CIIF} + \gamma_2 \pi_{t-1}^{CIIF} + e_{t+1} \tag{4.10}$$

$$\pi_{t+1}^{CIIF} = \alpha_r + \varphi_1 \pi_t^{CIIF} + \varphi_2 \pi_{t-1}^{CIIF} + v_{t+1}. \tag{4.11}$$

We calibrate these two processes to match in-sample estimates for monthly core inflation for a sample of OECD countries. Parameters:

$$\alpha_\pi = 0.15,\ \varphi_\pi = 0.90,\ \alpha_r = 0.05,\ \varphi_1 = 1.27,\ \varphi_2 = -0.3,\ \sigma_e^2 = 0.25^2,\ \sigma_v^2 = 0.1^2; \tag{4.12}$$
$$corr(e_{t+1},\ v_{t+1}) = 0.2;\ \text{under } H_0,\ \gamma_1 = \gamma_2 = 0;\ \text{under } H_A,\ \gamma_1 = 0.51;\ \gamma_2 = -0.50.$$

In contrast to our previous DGPs, DGP 3 is highly persistent in all three expressions (4.9), (4.10) and (4.11). We consider an initial estimation window of 240 observations ($R = 240$) and report results for several different number of predictions: $P = 120$, 180, 240 and 1000. Differing also from our previous DGPs, now we do not impose the correct number of lags for $\pi_t^{CIIF}$ in (4.10). We use BIC to choose the lag length with maximum lag $p = 6$, so in this DGP we deal with a certain degree of model uncertainty.

For each DGP we consider 5000 independent replications. In each replication, we generate 2000 observations on our dependent and independent variables. We discard the first 500 values to ensure stationarity. We evaluate the performance of our test and CW test using standard normal critical values at the 10%, 5% and 1% significance level for one sided tests. We construct estimates of the long run variance $\hat{V}$ in (2.8) using Newey and West (1987, 1994).

## 5. Simulation Results

In this section we present simulation results for size, size-adjusted-power and raw power of our tests. To save space, complete results are only reported when the nominal size is 5%.

Nevertheless, summary statistics for all three nominal sizes (10%, 5% and 1%) are also described in this section. Complete results for the nominal sizes of 10% and 1% are in Appendix A. We also report tables with the average across 5000 independent simulations of our *"power-booster-factor"*.

## 5.1 Simulation Results: Size

Results for nominal size 5% are in Table 1. From this table, in the rows labeled "CW" we see that the CW test is modestly undersized in all our DGPs and for all values of the number of forecasts $P$. This is also robust to the use of rolling and recursive windows. Table 6 indicates that the median size of CW is 0.037, below the nominal size of 0.05. Let us go back to Table 1. In the rows labeled "CW with PBF..." we present results for our approach. Three salient features are worth mentioning. First, empirical size is always higher than the equivalent figure for CW. Second, the empirical size of our approach is an increasing function of the parameter $\lambda$. Third; in Table 6 we see that the median size of our approach is 0.045, below the nominal size of 0.05. Nevertheless, our approach is not always undersized. In particular it is sometimes oversized for high values of $\lambda$. Notice, however, that in most entries of Table 1, results on size are better in our approach relative to CW. This means that size is higher than CW, but either below nominal size or slightly above it. This is particularly noticeable if we restrict ourselves to values of $\lambda$ equal or below 2. In this case it is only for DGP 2 and $P = 85$ that our approach is importantly oversized. This represents less than 10% of the relevant entries in Table 1. Other than that, our approach improves the empirical size of the CW test.

When the nominal size is 10%, the general picture described at the 5% level still stands. Table A.1 in Appendix A shows detailed results. CW still is modestly undersized, and our approach has always higher size than CW. Differing from the previous case (nominal size of 5%), now our approach is never heavily oversized, and most of the times it is slightly undersized. In the worst case we obtain an empirical size of 12.4%, which we consider tolerable. Table 6 reports the median size of CW and our approach. The corresponding figures are 7.3% for CW and 8.4% for our approach. In general terms, at the 10% level, our results are better relative to CW in terms of size.

A slightly different picture is shown for nominal sizes of 1%. CW is still slightly undersized, but now, in most entries of Table A.4 in Appendix A, our approach is slightly oversized with a median of 0.011 (see Table 6). In most cases size distortions with our approach are modest, but in some cases for large values of $\lambda$, our approach is importantly oversized. Interestingly, for low values of $\lambda$ ($1 \leq \lambda \leq 2$), the median size of our approach is 1%, and aside from the results of DGP 2 with $P = 85$, our test seems to be, in most cases, correctly sized.

## 5.2 Simulation Results: Power

Table 2 shows results for size-adjusted-power, Table 3 for power. Virtually all entries in Table 2 display higher size-adjusted-power for our approach relative to CW.[5] The only exception occurs

[5]Tables A.3 and A.6 in Appendix A show the same pattern when inference is carried out at the 10% and 1% significance levels.

**Table 1**

Empirical size: One-step-ahead forecasts, nominal size = 5%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.041 | 0.038 | 0.042 | 0.033 | 0.041 | 0.038 | 0.035 | 0.034 |
| CW with PBF $\lambda$ = 1.5 | 0.044 | 0.040 | 0.044 | 0.034 | 0.044 | 0.039 | 0.036 | 0.036 |
| CW with PBF $\lambda$ = 2.0 | 0.047 | 0.043 | 0.047 | 0.035 | 0.046 | 0.041 | 0.037 | 0.036 |
| CW with PBF $\lambda$ = 4.0 | 0.057 | 0.051 | 0.054 | 0.040 | 0.054 | 0.046 | 0.040 | 0.037 |
| CW with PBF $\lambda$ = 6.0 | 0.066 | 0.059 | 0.063 | 0.048 | 0.061 | 0.051 | 0.047 | 0.040 |
| CW | 0.036 | 0.033 | 0.037 | 0.031 | 0.038 | 0.034 | 0.032 | 0.031 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.055 | 0.046 | 0.044 | 0.047 | 0.053 | 0.044 | 0.042 | 0.035 |
| CW with PBF $\lambda$ = 1.5 | 0.058 | 0.049 | 0.048 | 0.049 | 0.057 | 0.046 | 0.043 | 0.037 |
| CW with PBF $\lambda$ = 2.0 | 0.060 | 0.052 | 0.050 | 0.051 | 0.060 | 0.047 | 0.045 | 0.038 |
| CW with PBF $\lambda$ = 4.0 | 0.072 | 0.063 | 0.059 | 0.057 | 0.072 | 0.058 | 0.049 | 0.041 |
| CW with PBF $\lambda$ = 6.0 | 0.083 | 0.073 | 0.069 | 0.061 | 0.080 | 0.064 | 0.055 | 0.043 |
| CW | 0.047 | 0.040 | 0.039 | 0.044 | 0.048 | 0.041 | 0.039 | 0.034 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.043 | 0.038 | 0.041 | 0.042 | 0.040 | 0.038 | 0.037 | 0.028 |
| CW with PBF $\lambda$ = 1.5 | 0.044 | 0.040 | 0.043 | 0.044 | 0.041 | 0.039 | 0.037 | 0.028 |
| CW with PBF $\lambda$ = 2.0 | 0.046 | 0.041 | 0.045 | 0.045 | 0.043 | 0.040 | 0.038 | 0.028 |
| CW with PBF $\lambda$ = 4.0 | 0.053 | 0.045 | 0.050 | 0.049 | 0.049 | 0.046 | 0.042 | 0.030 |
| CW with PBF $\lambda$ = 6.0 | 0.059 | 0.050 | 0.055 | 0.053 | 0.055 | 0.048 | 0.045 | 0.031 |
| CW | 0.038 | 0.036 | 0.038 | 0.041 | 0.038 | 0.036 | 0.035 | 0.027 |

*Notes*: 1. Table 1 displays empirical sizes for two tests of equal population mean squared prediction errors (MSPEs) against the one-sided alternative that one model has higher accuracy (lower MSPE). Rows with the label "CW" display results of the test proposed in Clark and West (2006, 2007). Rows with the label "CW with PBF..." display results of the test proposed in this paper. The term "PBF" stands for Power Booster Factor. Our test is the result of the multiplication of the t-statistic proposed in Clark and West (2006, 2007) and the Power Booster Factor presented in expression (2.17). This factor should be close to one under the null hypothesis, but should be greater than one under the alternative hypothesis. The implication is that our test should have more power than the test in Clark and West (2006, 2007). As the Power Booster Factor depends on the parameter $\lambda$, we present results for five different alternatives for this parameter: $\lambda = 1$; $\lambda = 1.5$; $\lambda = 2$; $\lambda = 4$; and $\lambda = 6$.

2. In DGP 1 the null model posits that the predictand $y_t$ is white noise, the alternative that $y_t$ depends on a constant and a variable $r_t$ that follows an autoregression of order 2. In DGPs 2 and 3, the null is that $y_t$ follows an AR(1), the alternative that $y_t$ is driven by additional variables following autoregressive processes. This implies that the univariate process for $y_t$ is not an AR(1). Section 4 of the main body of the paper gives exact specifications. In the exercises with DGP 1 and DGP 2 the alternative uses population lag lengths. In the exercises with DGP3 the alternative uses BIC to pick lags of the exogenous variable in the equation for $y_t$. All three DGPs are estimated by least squares.

3. Results are based on 5000 replications. A figure of 0.041 in the first column with numbers, for example, indicates that about 205 of the 5000 corresponding statistics were greater than 1.645, where 1.645 is the 5% critical value for a standard normal one-sided test.

4. Let $R$ be the rolling sample size (left panel in Table 1) or the smallest recursive sample used to estimate parameters needed under the alternative to make a forecast (right panel in Table 1). Then $R = 120$ in DGP 1 and DGP 2 and $R = 240$ in DGP 3. Table 1 shows results for several different numbers of predictions $P$. In DGP 1 we consider $P = 120$; $P = 240$; $P = 360$ and $P = 1000$. In DGP 2 we consider $P = 85$; $P = 170$; $P = 340$ and $P = 1000$. In DGP 3 we consider $P = 120$; $P = 180$; $P = 240$ and $P = 1000$. Results for nominal sizes of 10% and 1%, are available in the Appendix.

**Table 2**

Size-Adjusted-Power: One-step-ahead forecasts, nominal size = 5%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.372 | 0.506 | 0.604 | 0.901 | 0.429 | 0.622 | 0.765 | 0.986 |
| CW with PBF $\lambda$ = 1.5 | 0.372 | 0.506 | 0.605 | 0.901 | 0.434 | 0.624 | 0.766 | 0.986 |
| CW with PBF $\lambda$ = 2.0 | 0.374 | 0.506 | 0.605 | 0.901 | 0.437 | 0.625 | 0.766 | 0.986 |
| CW with PBF $\lambda$ = 4.0 | 0.380 | 0.511 | 0.608 | 0.900 | 0.449 | 0.628 | 0.768 | 0.986 |
| CW with PBF $\lambda$ = 6.0 | 0.385 | 0.517 | 0.611 | 0.901 | 0.453 | 0.630 | 0.770 | 0.986 |
| CW | 0.365 | 0.504 | 0.604 | 0.900 | 0.424 | 0.618 | 0.764 | 0.986 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.167 | 0.223 | 0.311 | 0.525 | 0.186 | 0.279 | 0.421 | 0.791 |
| CW with PBF $\lambda$ = 1.5 | 0.167 | 0.225 | 0.312 | 0.525 | 0.187 | 0.280 | 0.422 | 0.791 |
| CW with PBF $\lambda$ = 2.0 | 0.168 | 0.224 | 0.312 | 0.526 | 0.189 | 0.280 | 0.421 | 0.791 |
| CW with PBF $\lambda$ = 4.0 | 0.172 | 0.227 | 0.312 | 0.528 | 0.195 | 0.285 | 0.424 | 0.791 |
| CW with PBF $\lambda$ = 6.0 | 0.172 | 0.229 | 0.312 | 0.529 | 0.195 | 0.286 | 0.428 | 0.791 |
| CW | 0.167 | 0.223 | 0.313 | 0.523 | 0.185 | 0.274 | 0.419 | 0.790 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.666 | 0.763 | 0.804 | 0.993 | 0.747 | 0.861 | 0.917 | 1.000 |
| CW with PBF $\lambda$ = 1.5 | 0.667 | 0.766 | 0.806 | 0.993 | 0.752 | 0.863 | 0.918 | 1.000 |
| CW with PBF $\lambda$ = 2.0 | 0.671 | 0.768 | 0.807 | 0.993 | 0.756 | 0.864 | 0.918 | 1.000 |
| CW with PBF $\lambda$ = 4.0 | 0.688 | 0.773 | 0.811 | 0.993 | 0.773 | 0.870 | 0.921 | 1.000 |
| CW with PBF $\lambda$ = 6.0 | 0.696 | 0.779 | 0.813 | 0.993 | 0.780 | 0.874 | 0.921 | 1.000 |
| CW | 0.657 | 0.760 | 0.802 | 0.993 | 0.738 | 0.856 | 0.913 | 1.000 |

*Notes*: 1. Table 2 displays figures on size-adjusted-power for two tests of equal MSPEs against the one-sided alternative that one model has higher accuracy (lower MSPE). Size-adjusted-power represents the percentage of correct rejections of the null hypothesis enforcing the empirical size of the tests to coincide with their nominal size. Rows with the label "CW" display results of the test proposed in Clark and West (2006, 2007). Rows with the label "CW with PBF..." display results of the test proposed in this paper. The term "PBF" stands for Power Booster Factor. Our test is the result of the multiplication of the t-statistic proposed in Clark and West (2006, 2007) and the Power Booster Factor presented in expression (2.17). This factor should be close to one under the null hypothesis, but should be greater than one under the alternative hypothesis. The implication is that our test should have more power than the test in Clark and West (2006, 2007). As the Power Booster Factor depends on the parameter $\lambda$, we present results for five different alternatives for this parameter: $\lambda = 1$; $\lambda = 1.5$; $\lambda = 2$; $\lambda = 4$; and $\lambda = 6$.
2. See notes to Table 1 for further details.

for DGP 2, when $P = 340$ under the rolling scheme. In all other cases size-adjusted-power is higher when using our *"power-booster-factor"*. Differences are in general low, however. Table 6 shows the median size-adjusted-power for all three nominal sizes. The figures for CW are 0.741; 0.638 and 0.401 when nominal sizes are 10%, 5% and 1% respectively. The equivalent figures of our new approach are 0.745; 0.648 and 0.431. Consistent with our beliefs, gains in size-adjusted-power relative to CW are tiny when inference is carried out at the 10% level, small to moderate at the 5% level, and substantial when inference is carried out at the 1% level. Table 6 shows median results, but also consistent with our beliefs, Table 2 shows that size-adjusted-power is an increasing function of $\lambda$, so gains relative to CW are more important when $\lambda$ is high. To give an example, Table 2 indicates that for DGP 3, under the rolling scheme, when $P = 120$, CW has

a figure of size-adjusted-power equal to 0.657. For $\lambda = 1$, the equivalent figure of our approach is 0.666, only a tiny improvement relative to CW. Nevertheless, for $\lambda = 6$, our figure is 0.696, a considerable gain relative to CW. The same gains are less impressive at the 10% nominal size (see Table A.2 in Appendix A) but much more impressive when nominal size is 1%. In this case, the equivalent entries in Table A.5 in Appendix A show a figure for CW of 0.441. For our approach when $\lambda = 6$, our figure is 0.558. (See in Table A.5 the case of DGP 3 under the rolling scheme, when $P = 120$). A final point: gains in size-adjusted-power are more important for small and moderate values of the number of predictions $P$. Asymptotically, gains in size-adjusted-power tend to disappear.

Table 3 shows result on raw power. Virtually all entries in Table 3 display higher power for

**Table 3**

Raw Power: One-step-ahead forecasts, nominal size = 5%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R = 120$) | | | | DGP 1 ($R = 120$) | | | |
| Test | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ |
| CW with PBF $\lambda = 1.0$ | 0.335 | 0.453 | 0.572 | 0.875 | 0.399 | 0.567 | 0.714 | 0.977 |
| CW with PBF $\lambda = 1.5$ | 0.347 | 0.465 | 0.583 | 0.877 | 0.411 | 0.575 | 0.722 | 0.977 |
| CW with PBF $\lambda = 2.0$ | 0.363 | 0.477 | 0.592 | 0.880 | 0.420 | 0.583 | 0.726 | 0.977 |
| CW with PBF $\lambda = 4.0$ | 0.402 | 0.511 | 0.624 | 0.889 | 0.460 | 0.611 | 0.745 | 0.979 |
| CW with PBF $\lambda = 6.0$ | 0.431 | 0.537 | 0.646 | 0.897 | 0.488 | 0.634 | 0.756 | 0.981 |
| CW | 0.303 | 0.428 | 0.550 | 0.868 | 0.368 | 0.548 | 0.703 | 0.976 |
| | DGP 2 ($R = 120$) | | | | DGP 2 ($R = 120$) | | | |
| | $P = 85$ | $P = 170$ | $P = 340$ | $P = 1000$ | $P = 85$ | $P = 170$ | $P = 340$ | $P = 1000$ |
| CW with PBF $\lambda = 1.0$ | 0.180 | 0.212 | 0.296 | 0.517 | 0.197 | 0.259 | 0.389 | 0.759 |
| CW with PBF $\lambda = 1.5$ | 0.186 | 0.220 | 0.304 | 0.523 | 0.205 | 0.265 | 0.395 | 0.762 |
| CW with PBF $\lambda = 2.0$ | 0.195 | 0.228 | 0.311 | 0.529 | 0.211 | 0.271 | 0.400 | 0.765 |
| CW with PBF $\lambda = 4.0$ | 0.224 | 0.261 | 0.340 | 0.550 | 0.241 | 0.300 | 0.422 | 0.771 |
| CW with PBF $\lambda = 6.0$ | 0.248 | 0.285 | 0.360 | 0.564 | 0.266 | 0.320 | 0.439 | 0.777 |
| CW | 0.159 | 0.192 | 0.281 | 0.507 | 0.177 | 0.240 | 0.372 | 0.754 |
| | DGP 3 ($R = 240$) | | | | DGP 3 ($R = 240$) | | | |
| | $P = 120$ | $P = 180$ | $P = 240$ | $P = 1000$ | $P = 120$ | $P = 180$ | $P = 240$ | $P = 1000$ |
| CW with PBF $\lambda = 1.0$ | 0.645 | 0.734 | 0.786 | 0.992 | 0.721 | 0.836 | 0.897 | 1.000 |
| CW with PBF $\lambda = 1.5$ | 0.656 | 0.741 | 0.790 | 0.993 | 0.731 | 0.842 | 0.900 | 1.000 |
| CW with PBF $\lambda = 2.0$ | 0.662 | 0.746 | 0.794 | 0.993 | 0.740 | 0.846 | 0.903 | 1.000 |
| CW with PBF $\lambda = 4.0$ | 0.696 | 0.763 | 0.810 | 0.993 | 0.770 | 0.862 | 0.912 | 1.000 |
| CW with PBF $\lambda = 6.0$ | 0.719 | 0.777 | 0.822 | 0.994 | 0.789 | 0.873 | 0.919 | 1.000 |
| CW | 0.622 | 0.716 | 0.776 | 0.992 | 0.697 | 0.825 | 0.890 | 1.000 |

*Notes*: 1. Table 3 displays figures on power (also called raw power) for two tests of equal population MSPEs against the one-sided alternative that one model has higher accuracy (lower MSPE). Power represents the percentage of correct rejections of the null hypothesis using standard normal critical values. Rows with the label "CW" display results of the test proposed in Clark and West (2006, 2007). Rows with the label "CW with PBF..." display results of the test proposed in this paper. The term "PBF" stands for Power Booster Factor. Our test is the result of the multiplication of the t-statistic proposed in Clark and West (2006, 2007) and the Power Booster Factor presented in expression (2.17). This factor should be close to one under the null hypothesis, but should be greater than one under the alternative hypothesis. The implication is that our test should have more power than the test in Clark and West (2006, 2007). As the Power Booster Factor depends on the parameter $\lambda$, we present results for five different alternatives for this parameter: $\lambda = 1$; $\lambda = 1.5$; $\lambda = 2$; $\lambda = 4$; and $\lambda = 6$.

2. See notes to Table 1 for further details.

our approach relative to CW.[6] The only exception occurs for DGP 3, when $P = 1000$ under the recursive scheme. In this case figures on power are all equal to one both in CW and our approach. In all other cases power is higher when using our *"power-booster-factor"*. Differing from our previous analysis on size-adjusted-power, now the gains of our approach relative to CW are, in general, substantial. Table 6 shows the median power for all three nominal sizes. The figures for CW are 0.702; 0.586 and 0.342 when nominal sizes are 10%, 5% and 1% respectively. The equivalent figures of our new approach are 0.731; 0.646 and 0.468. Notice that for CW median figures on power are lower than on size-adjusted-power, reflecting the fact that CW is a little undersized. When using our *"power-booster-factor"* we find mixed results. Figures on power are lower than on size-adjusted-power when inference is carried out at the 10% significance level, which is consistent with our approach being a little undersized. Nevertheless, figures on power are higher than on size-adjusted-power when inference is carried out at the 5% and 1% significance levels, which is consistent with our approach being a little oversized, especially for high values of $\lambda$, when the nominal size is set to 1%.

Gains in power relative to CW are moderate when inference is carried out at the 10% level, higher at the 5% level, and huge when inference is carried out at the 1% level. Table 3 also shows that power is an increasing function of $\lambda$, so gains relative to CW are more important when $\lambda$ is high. To give an example, Table 3 indicates that for DGP 1, under the recursive scheme, when $P = 120$, CW has a figure on power equal to 0.368. For $\lambda = 1$, the equivalent figure of our approach is 0.399, a small improvement relative to CW. Nevertheless, for $\lambda = 6$, our figure is 0.488, a substantial gain relative to CW. The same gains are slightly lower at the 10% nominal size (see Table A.3 in Appendix A) but much more impressive when nominal size is 1%. In this case, the equivalent entries in Table A.6 in Appendix A show a figure for CW of 0.147. For our approach when $\lambda = 6$, our figure is 0.343. (See in Table A.3 the case of DGP 1 under the recursive scheme, when $P = 120$). Gains in power are more important for small and moderate values of the number of predictions $P$ and tend to disappear as the number of predictions grows to infinity.

Notice that some of the gains in power come from comparing a slightly oversized test with a slightly undersized test. For instance, in rolling regressions for DGP 1, $P = 360$ and $\lambda = 4$, Table 1 shows a figure on size for our approach of 0.054. The same figure for CW is 0.037. Gains in power in this case are high. Table 3 shows that our approach has raw power equal to 62.4%. The equivalent figure for CW is 55%. Not all the gains in power come from comparing undersized to oversized tests. In many cases our approach generates a less undersized test than CW. This, plus some gains in size-adjusted-power, generates a test with higher raw power.

Based upon our simulation results we see that both size and power are increasing functions of $\lambda$. To avoid the risk of an oversized test, we think that an adequate recommendation for empirical work is the following: For inference at the 10% significance level set $\lambda$ to 4. For inference at the 5% significance level set $\lambda$ to 2, and for inference at the 1% significance level set $\lambda$ to 1. This recommendation is based on the observation that the risk of obtaining an oversized test is higher

[6]Tables A.2 and A.5 in Appendix A show the same pattern when inference is carried out at the 10% and 1% significance levels.

at tighter significance levels. A more general and conservative approach would suggest to use $\lambda = 1$ irrespective of the significance level.

A final point: Tables 4 and 5 shows the average across our 5000 simulations of our *"power-booster-factor"* computed both under the null and alternative hypotheses. As expected, our factor is very close to one under the null, and greater than 1 under the alternative hypothesis. These figures are consistent with our results shown in previous tables, both on size and power.

**Table 4**

*Power-Booster-Factor* under the null hypothesis: One-step-ahead forecasts.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R = 120$) | | | | DGP 1 ($R = 120$) | | | |
| Test | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ |
| $\lambda = 1.0$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 1.5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 2.0$ | 1.000 | 1.000 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 4.0$ | 1.003 | 1.002 | 1.002 | 1.001 | 1.002 | 1.001 | 1.001 | 1.000 |
| $\lambda = 6.0$ | 1.008 | 1.005 | 1.005 | 1.002 | 1.005 | 1.003 | 1.002 | 1.001 |
| | DGP 2 ($R = 120$) | | | | DGP 2 ($R = 120$) | | | |
| | $P = 85$ | $P = 170$ | $P = 340$ | $P = 1000$ | $P = 85$ | $P = 170$ | $P = 340$ | $P = 1000$ |
| $\lambda = 1.0$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 1.5$ | 1.000 | 1.000 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 2.0$ | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 4.0$ | 1.005 | 1.004 | 1.003 | 1.002 | 1.003 | 1.001 | 1.001 | 1.004 |
| $\lambda = 6.0$ | 1.014 | 1.009 | 1.006 | 1.006 | 1.008 | 1.004 | 1.002 | 1.001 |
| | DGP 3 ($R = 240$) | | | | DGP 3 ($R = 240$) | | | |
| | $P = 120$ | $P = 180$ | $P = 240$ | $P = 1000$ | $P = 120$ | $P = 180$ | $P = 240$ | $P = 1000$ |
| $\lambda = 1.0$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 1.5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 2.0$ | 1.000 | 1.000 | 1.001 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $\lambda = 4.0$ | 1.002 | 1.001 | 1.002 | 1.001 | 1.001 | 1.001 | 1.001 | 1.000 |
| $\lambda = 6.0$ | 1.004 | 1.003 | 1.003 | 1.001 | 1.003 | 1.002 | 1.002 | 1.001 |

*Notes*: 1. Table 4 displays the average of our *Power-Booster-Factor* presented in expression (2.17) across our 5000 replications when the null hypothesis is true in our three DGPs. This factor should be close to one under the null hypothesis, but should be greater than one under the alternative hypothesis.
2. See notes to Table 1 for further details.

**Table 5**

*Power-Booster-Factor* under the alternative hypothesis: One-step-ahead forecasts.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R = 120$) | | | | DGP 1 ($R = 120$) | | | |
| Test | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ | $P = 120$ | $P = 240$ | $P = 360$ | $P = 1000$ |
| $\lambda = 1.0$ | 1.040 | 1.040 | 1.041 | 1.041 | 1.040 | 1.041 | 1.040 | 1.041 |
| $\lambda = 1.5$ | 1.062 | 1.062 | 1.062 | 1.062 | 1.062 | 1.062 | 1.062 | 1.062 |
| $\lambda = 2.0$ | 1.084 | 1.084 | 1.084 | 1.084 | 1.084 | 1.084 | 1.083 | 1.084 |
| $\lambda = 4.0$ | 1.181 | 1.181 | 1.179 | 1.177 | 1.184 | 1.180 | 1.177 | 1.176 |
| $\lambda = 6.0$ | 1.310 | 1.294 | 1.287 | 1.280 | 1.304 | 1.291 | 1.284 | 1.279 |

(Continued on next page)

**Table 5** (continued)

| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
|---|---|---|---|---|---|---|---|---|
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| $\lambda$ = 1.0 | 1.021 | 1.021 | 1.022 | 1.022 | 1.021 | 1.021 | 1.022 | 1.022 |
| $\lambda$ = 1.5 | 1.032 | 1.033 | 1.033 | 1.033 | 1.032 | 1.032 | 1.033 | 1.032 |
| $\lambda$ = 2.0 | 1.044 | 1.044 | 1.045 | 1.044 | 1.044 | 1.044 | 1.044 | 1.044 |
| $\lambda$ = 4.0 | 1.098 | 1.096 | 1.094 | 1.091 | 1.096 | 1.093 | 1.092 | 1.090 |
| $\lambda$ = 6.0 | 1.163 | 1.155 | 1.149 | 1.142 | 1.156 | 1.148 | 1.145 | 1.139 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| $\lambda$ = 1.0 | 1.077 | 1.078 | 1.078 | 1.079 | 1.081 | 1.082 | 1.083 | 1.086 |
| $\lambda$ = 1.5 | 1.119 | 1.120 | 1.120 | 1.121 | 1.125 | 1.127 | 1.128 | 1.131 |
| $\lambda$ = 2.0 | 1.164 | 1.164 | 1.164 | 1.165 | 1.172 | 1.174 | 1.175 | 1.179 |
| $\lambda$ = 4.0 | 1.370 | 1.368 | 1.366 | 1.361 | 1.386 | 1.388 | 1.389 | 1.394 |
| $\lambda$ = 6.0 | 1.633 | 1.624 | 1.618 | 1.595 | 1.656 | 1.653 | 1.652 | 1.651 |

*Notes*: 1. Table 5 displays the average of our *Power-Booster-Factor* presented in expression (2.17) across our 5000 replications when the alternative hypothesis is true in our three DGPs. This factor should be close to one under the null hypothesis, but should be greater than one under the alternative hypothesis.
2. See notes to Table 1 for further details.

**Table 6**
Summary statistics from Monte Carlo simulations.

| | Median Empirical Size | | |
|---|---|---|---|
| Tests | nominal size is 10% | nominal size is 5% | nominal size is 1% |
| CW with PBF | 0.084 | 0.045 | 0.011 |
| CW | 0.073 | 0.037 | 0.007 |
| | Median Size-Adjusted-Power | | |
| | nominal size is 10% | nominal size is 5% | nominal size is 1% |
| CW with PBF | 0.745 | 0.648 | 0.431 |
| CW | 0.741 | 0.638 | 0.401 |
| | Median Power | | |
| | nominal size is 10% | nominal size is 5% | nominal size is 1% |
| CW with PBF | 0.731 | 0.646 | 0.468 |
| CW | 0.702 | 0.586 | 0.342 |

*Notes*: 1. Table 6 displays the median across 5000 replications of figures on size, size-adjusted-power and power for the test in Clark and West (2006, 2007) and the test proposed in this paper. We present results for three nominal sizes: 10%, 5% and 1%.
2. In Table 6 "CW" stands for "Clark and West", whereas "CW with PBF" stands for "Clark and West with Power Booster Factor" which corresponds to our contribution. As the *"power-booster-factor"* depends on the parameter $\lambda$ (see 2.17) the median is taken across all the five values of $\lambda$ we consider in our simulations.
3. See notes to Table 1 for further details.

## 6. Empirical Illustration

We consider predicting core domestic inflation with an international core factor. As mentioned in Section 4, recent literature has explored the predictive linkages between domestic and international inflation concluding that this linkage is important both at the core and headline level

for several countries. See for instance Ciccarelli and Mojon (2010), Duncan and Martínez-García (2015), Kabukcuoglu and Martínez-García (2018), Morales-Arias and Moura (2013), Hakkio (2009), Pincheira and Gatty (2016) and Medel et al. (2016).

We consider nested models similar, but not equal, to those in (4.9) and (4.10).

Let $\pi_{it}$ be year-on-year domestic core inflation rates in country $i$. Following the literature cited in the previous paragraph, we build a core international inflation factor (CIIF) as the simple average of $\pi_{it}$ measured using monthly core CPI data, with $i$ ranging over 30 OECD countries:[7]

$$\pi_t^{CIIF} = \frac{1}{30}\sum_{(i=1)}^{30} \pi_{it}. \tag{6.1}$$

Our data range from January 1995 to December 2015 (252 observations). We focus on core inflation measured as CPI inflation excluding food and energy components. For the out-of-sample analysis we estimate our models by OLS in recursive windows with an initial window length of 100 observations ($R = 100$, from January 1995 to April 2003). This means that our first one-step-ahead forecast is made for May 2003, while the last one is made for December 2015. We focus only on one-step-ahead forecasts. We analyze if the CIIF has the ability to predict inflation for all the 30 OECD countries included in the average in (6.1). For each country, we consider the following nested models:

$$\text{(model 1: null model)}: \qquad \pi_{(it+1)} = \alpha_\pi + \beta_1\pi_{it} + \beta_2\pi_{(it-11)} + e_{(t+1)} \tag{6.2}$$

$$\text{(model 2: alternative model)}: \qquad \pi_{(it+1)} = \alpha_\pi + \beta_1\pi_{it} + \beta_2\pi_{(it-11)} + \gamma(B)\pi_t^{CIIF} + e_{(t+1)} \tag{6.3}$$

Here, $\gamma(B) = \sum_{(j=0)}^{q} \gamma_j B^j$ represents a lag polynomial and B is the backshift operator such that $B^j X_t = X_{(t-j)}$. The lag order $q$ is selected in each estimation window with BIC with $1 \leq q \leq 12$. Notice that for this empirical illustration we consider 6 different values for the parameter $\lambda$ implicitly defined in expression (2.17). We consider $\lambda = 0$, which is nothing but the CW t-statistic, plus the following values: $\lambda = 1$; 1.5; 2; 4 and 6. Table 7 shows summary results. In particular, this table shows the percentage of countries for which each test rejects the null hypothesis. We present results at the three usual significance levels: 10%, 5% and 1%.

Consistent with our simulations, CW tends to reject less frequently than our new approach. This is uniform across different nominal sizes. The highest difference between CW and our test occurs when nominal size is 1%. Here CW rejects the null in only 4 countries (13.3%) but, depending on our preferred value for $\lambda$, our approach rejects in a range of 6 to 10 countries (20% to 33.3%). Now, simulation evidence included in Section 5 suggests that for high values of the parameter $\lambda$ our test might be oversized when using a nominal size of 1%. Nevertheless, if we look at our results at the 10% level we still obtain more rejections with our new approach, even when using a moderate choice for $\lambda$. For instance, when using $\lambda = 2$ and a nominal size of 10%, our approach rejects the null of no predictability in 33.3% of the countries whereas the equivalent

[7] We consider the following countries: Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, The Netherlands, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, U.K. and the U.S. (Data source: OECD Main Economic Indicators.)

figure using the widely used CW test is only 23.3%.[8] Let us recall that our simulations show an adequate empirical size for our approach when the nominal size is 10%.

**Table 7**

Share of OECD countries for which the null of no predictability is rejected. The null posits that an international core inflation factor does not predict domestic core inflation.

| | $\lambda = 0$ (CW) | $\lambda = 1$ | $\lambda = 1.5$ | $\lambda = 2$ | $\lambda = 4$ | $\lambda = 6$ |
|---|---|---|---|---|---|---|
| Rejection at the 10% | | | | | | |
| CW with *Power-Booster-Factor* | 0.233 | 0.300 | 0.300 | 0.333 | 0.333 | 0.333 |
| Rejection at the 5% | | | | | | |
| CW with *Power-Booster-Factor* | 0.233 | 0.233 | 0.233 | 0.233 | 0.300 | 0.300 |
| Rejection at the 1% | | | | | | |
| CW with *Power-Booster-Factor* | 0.133 | 0.200 | 0.233 | 0.233 | 0.300 | 0.333 |

*Notes*: 1. In this table forecasts from an autoregression (null model, or model 1) for year-on-year monthly core CPI inflation rate are compared to forecasts coming from an alternative model (model 2) that augments model 1 with a measure of international core inflation. See (6.1), (6.2) and (6.3) for details.
2. International core inflation is defined as the simple average of monthly year-on-year domestic core CPI inflation rates for 30 OECD economies. Core CPI is defined as CPI excluding food and energy components.
3. Table 7 shows the share of the total of 30 OECD countries in our sample for which our tests reject the null hypothesis. In the column with the label "$\lambda = 0$ (CW)" our test coincides with the test proposed in Clark and West (2006, 2007), so rejections rates in that column corresponds to rejection rates of the test in Clark and West (2006, 2007). In the rest of the columns, Table 7 presents rejection rates of our test for different values of the parameter $\lambda$. (See (2.17)). When $\lambda > 0$ our test is different to the test in Clark and West (2006, 2007).
4. Data are described in the text. See notes to earlier tables for additional definitions.

## 7. Summary and Concluding Remarks

In this paper we introduce a *"power-booster-factor"* for out-of-sample tests of predictability. This factor can be used to improve finite sample properties of several out-of-sample tests of predictability. Yet, in this paper, we focus on the widely used test developed by Clark and West (2006, 2007). We construct a new standard normal test multiplying the CW t-statistic by our *"power-booster-factor"*. The key idea relies on the fact that this factor should be close to one under the null hypothesis of no predictability, but should be greater than one under the alternative hypothesis.

Monte Carlo simulations reveal that our new test is, generally speaking, well sized and powerful. In particular, it is less undersized, more powerful and sometimes much more powerful than the test by Clark and West (2006, 2007). We notice, however, that improvements in size-adjusted-power are moderate, and gains in power are mostly induced by our test being less undersized than CW.

[8] CW rejects the null of no predictability for Czech Republic, Iceland, Korea, Luxembourg, Portugal, Slovak Republic and Turkey, whereas our approach rejects for the same countries plus Chile, Israel and the U.S.

Both size and power of our new approach are increasing functions of a parameter that we have denoted by $\lambda$, and that the researcher needs to pick in advance. Although we have not developed a theory yet on how to pick this parameter, our Monte Carlo simulations shed some light in this regard. To avoid the risk of an oversized test, we think that an adequate recommendation for empirical work is the following: For inference at the 10% significance level set $\lambda$ to 4. For inference at the 5% significance level set $\lambda$ to 2, and for inference at the 1% significance level set $\lambda$ to 1. This recommendation is based on the observation that the risk of obtaining an oversized test is higher at tighter significance levels. A more general and conservative approach would suggest to use $\lambda = 1$ irrespective of the significance level.

To illustrate the use of our test we present an empirical application in the context of inflation forecasts. Based on a vast literature exploring the predictive linkages between domestic and international inflation, we analyze the predictive ability of an international core inflation factor to forecast domestic core inflation. We consider the case of 30 OECD economies with monthly observations for the period January 1995-March 2015. Consistent with the structure of our test and with our simulations, CW tends to reject less frequently than our new approach. This is uniform across different nominal sizes. For instance, our approach rejects the null of no predictability in at least 30.0% of the countries when the nominal size is 10%. The equivalent figure using the widely used CW test is only 23.3%. At the 1% level, our test rejects the null in at least 20% of the countries whereas the CW test rejects only in 13.3% of the countries. Our results suggest a strong influence of global inflation in a selected group of our sample of OECD countries.

A natural extension for further research could explore in more detail how to pick our $\lambda$ parameter, could also explore the application of variants of our factor to improve the finite sample behavior of other testing strategies, and could also evaluate the performance of our test when the focus of interest are multistep ahead forecasts.

## Appendix A - Tables

### Table A.1

Empirical size: One-step-ahead forecasts, nominal size = 10%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.078 | 0.075 | 0.085 | 0.082 | 0.078 | 0.074 | 0.069 | 0.065 |
| CW with PBF $\lambda$ = 1.5 | 0.080 | 0.078 | 0.086 | 0.082 | 0.080 | 0.077 | 0.071 | 0.065 |
| CW with PBF $\lambda$ = 2.0 | 0.084 | 0.081 | 0.088 | 0.084 | 0.081 | 0.078 | 0.072 | 0.066 |
| CW with PBF $\lambda$ = 4.0 | 0.096 | 0.091 | 0.096 | 0.089 | 0.090 | 0.084 | 0.077 | 0.068 |
| CW with PBF $\lambda$ = 6.0 | 0.106 | 0.100 | 0.103 | 0.093 | 0.097 | 0.090 | 0.080 | 0.070 |
| CW | 0.073 | 0.072 | 0.080 | 0.079 | 0.074 | 0.072 | 0.065 | 0.064 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.092 | 0.086 | 0.090 | 0.097 | 0.091 | 0.083 | 0.077 | 0.073 |
| CW with PBF $\lambda$ = 1.5 | 0.095 | 0.090 | 0.093 | 0.099 | 0.093 | 0.086 | 0.784 | 0.074 |
| CW with PBF $\lambda$ = 2.0 | 0.097 | 0.093 | 0.095 | 0.100 | 0.096 | 0.088 | 0.079 | 0.075 |
| CW with PBF $\lambda$ = 4.0 | 0.111 | 0.104 | 0.103 | 0.106 | 0.105 | 0.095 | 0.085 | 0.077 |
| CW with PBF $\lambda$ = 6.0 | 0.124 | 0.114 | 0.110 | 0.114 | 0.115 | 0.102 | 0.091 | 0.079 |
| CW | 0.084 | 0.080 | 0.086 | 0.095 | 0.086 | 0.079 | 0.075 | 0.072 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.083 | 0.074 | 0.074 | 0.083 | 0.077 | 0.071 | 0.072 | 0.058 |
| CW with PBF $\lambda$ = 1.5 | 0.084 | 0.077 | 0.075 | 0.084 | 0.078 | 0.073 | 0.072 | 0.059 |
| CW with PBF $\lambda$ = 2.0 | 0.087 | 0.080 | 0.076 | 0.084 | 0.079 | 0.073 | 0.073 | 0.059 |
| CW with PBF $\lambda$ = 4.0 | 0.094 | 0.085 | 0.082 | 0.087 | 0.085 | 0.078 | 0.077 | 0.061 |
| CW with PBF $\lambda$ = 6.0 | 0.099 | 0.093 | 0.088 | 0.089 | 0.090 | 0.082 | 0.080 | 0.062 |
| CW | 0.077 | 0.072 | 0.071 | 0.083 | 0.073 | 0.068 | 0.069 | 0.058 |

*Notes*: 1. Table A.1 is equivalent to Table 1 in the main body of the paper with the only difference that the nominal size in Table A.1 is 10% and not 5% as in Table 1.
2. See notes to Table 1 for further details.

**Table A.2**

Size-Adjusted-Power: One-step-ahead forecasts, nominal size = 10%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.501 | 0.626 | 0.720 | 0.940 | 0.577 | 0.735 | 0.851 | 0.995 |
| CW with PBF $\lambda$ = 1.5 | 0.501 | 0.626 | 0.720 | 0.940 | 0.578 | 0.735 | 0.851 | 0.995 |
| CW with PBF $\lambda$ = 2.0 | 0.500 | 0.626 | 0.721 | 0.940 | 0.577 | 0.736 | 0.851 | 0.995 |
| CW with PBF $\lambda$ = 4.0 | 0.501 | 0.628 | 0.721 | 0.939 | 0.581 | 0.739 | 0.851 | 0.995 |
| CW with PBF $\lambda$ = 6.0 | 0.503 | 0.627 | 0.725 | 0.940 | 0.584 | 0.739 | 0.851 | 0.995 |
| CW | 0.501 | 0.623 | 0.719 | 0.939 | 0.575 | 0.733 | 0.850 | 0.995 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.278 | 0.335 | 0.426 | 0.641 | 0.302 | 0.396 | 0.556 | 0.872 |
| CW with PBF $\lambda$ = 1.5 | 0.279 | 0.336 | 0.427 | 0.641 | 0.303 | 0.396 | 0.556 | 0.872 |
| CW with PBF $\lambda$ = 2.0 | 0.279 | 0.336 | 0.428 | 0.641 | 0.303 | 0.397 | 0.557 | 0.872 |
| CW with PBF $\lambda$ = 4.0 | 0.277 | 0.337 | 0.430 | 0.640 | 0.306 | 0.399 | 0.559 | 0.872 |
| CW with PBF $\lambda$ = 6.0 | 0.278 | 0.338 | 0.430 | 0.640 | 0.306 | 0.401 | 0.558 | 0.872 |
| CW | 0.277 | 0.335 | 0.426 | 0.640 | 0.301 | 0.395 | 0.556 | 0.872 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.750 | 0.821 | 0.862 | 0.996 | 0.832 | 0.908 | 0.948 | 1.000 |
| CW with PBF $\lambda$ = 1.5 | 0.752 | 0.822 | 0.863 | 0.996 | 0.834 | 0.909 | 0.949 | 1.000 |
| CW with PBF $\lambda$ = 2.0 | 0.753 | 0.824 | 0.863 | 0.996 | 0.835 | 0.909 | 0.949 | 1.000 |
| CW with PBF $\lambda$ = 4.0 | 0.758 | 0.828 | 0.864 | 0.996 | 0.842 | 0.910 | 0.950 | 1.000 |
| CW with PBF $\lambda$ = 6.0 | 0.764 | 0.830 | 0.865 | 0.996 | 0.846 | 0.912 | 0.950 | 1.000 |
| CW | 0.748 | 0.818 | 0.862 | 0.996 | 0.829 | 0.906 | 0.948 | 1.000 |

*Notes*: 1. Table A.2 is equivalent to Table 2 in the main body of the paper with the only difference that the nominal size in Table A.2 is 10% and not 5% as in Table 2.
2. See notes to Table 1 for further details.

**Table A.3**

Raw Power: One-step-ahead forecasts, nominal size = 10%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.457 | 0.579 | 0.691 | 0.926 | 0.519 | 0.687 | 0.805 | 0.989 |
| CW with PBF $\lambda$ = 1.5 | 0.464 | 0.585 | 0.694 | 0.926 | 0.527 | 0.692 | 0.807 | 0.989 |
| CW with PBF $\lambda$ = 2.0 | 0.470 | 0.591 | 0.697 | 0.927 | 0.535 | 0.695 | 0.807 | 0.989 |
| CW with PBF $\lambda$ = 4.0 | 0.493 | 0.610 | 0.716 | 0.931 | 0.557 | 0.707 | 0.821 | 0.990 |
| CW with PBF $\lambda$ = 6.0 | 0.515 | 0.626 | 0.730 | 0.935 | 0.576 | 0.720 | 0.829 | 0.991 |
| CW | 0.437 | 0.565 | 0.683 | 0.924 | 0.502 | 0.677 | 0.800 | 0.989 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.260 | 0.308 | 0.409 | 0.634 | 0.282 | 0.357 | 0.503 | 0.836 |
| CW with PBF $\lambda$ = 1.5 | 0.266 | 0.317 | 0.412 | 0.637 | 0.288 | 0.363 | 0.506 | 0.838 |
| CW with PBF $\lambda$ = 2.0 | 0.272 | 0.324 | 0.417 | 0.639 | 0.294 | 0.367 | 0.509 | 0.839 |
| CW with PBF $\lambda$ = 4.0 | 0.294 | 0.344 | 0.433 | 0.650 | 0.316 | 0.386 | 0.525 | 0.844 |
| CW with PBF $\lambda$ = 6.0 | 0.314 | 0.359 | 0.446 | 0.660 | 0.333 | 0.404 | 0.537 | 0.847 |
| CW | 0.245 | 0.294 | 0.399 | 0.626 | 0.269 | 0.345 | 0.493 | 0.834 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.731 | 0.798 | 0.841 | 0.995 | 0.804 | 0.886 | 0.932 | 1.000 |
| CW with PBF $\lambda$ = 1.5 | 0.734 | 0.802 | 0.844 | 0.995 | 0.809 | 0.888 | 0.932 | 1.000 |
| CW with PBF $\lambda$ = 2.0 | 0.738 | 0.804 | 0.847 | 0.995 | 0.813 | 0.892 | 0.935 | 1.000 |
| CW with PBF $\lambda$ = 4.0 | 0.752 | 0.812 | 0.853 | 0.995 | 0.826 | 0.898 | 0.939 | 1.000 |
| CW with PBF $\lambda$ = 6.0 | 0.762 | 0.823 | 0.859 | 0.996 | 0.836 | 0.904 | 0.942 | 1.000 |
| CW | 0.722 | 0.789 | 0.838 | 0.995 | 0.793 | 0.882 | 0.929 | 1.000 |

*Notes*: 1. Table A.3 is equivalent to Table 3 in the main body of the paper with the only difference that the nominal size in Table A.3 is 10% and not 5% as in Table 3.
2. See notes to Table 1 for further details.

**Table A.4**

Empirical Size: One-step-ahead forecasts, nominal size = 1%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.009 | 0.008 | 0.007 | 0.006 | 0.011 | 0.011 | 0.010 | 0.008 |
| CW with PBF $\lambda$ = 1.5 | 0.011 | 0.010 | 0.008 | 0.007 | 0.012 | 0.012 | 0.010 | 0.008 |
| CW with PBF $\lambda$ = 2.0 | 0.013 | 0.010 | 0.009 | 0.007 | 0.013 | 0.013 | 0.011 | 0.008 |
| CW with PBF $\lambda$ = 4.0 | 0.021 | 0.014 | 0.013 | 0.009 | 0.020 | 0.016 | 0.013 | 0.010 |
| CW with PBF $\lambda$ = 6.0 | 0.030 | 0.020 | 0.019 | 0.011 | 0.026 | 0.021 | 0.016 | 0.011 |
| CW | 0.007 | 0.005 | 0.005 | 0.005 | 0.010 | 0.009 | 0.008 | 0.007 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.015 | 0.010 | 0.009 | 0.009 | 0.016 | 0.011 | 0.010 | 0.007 |
| CW with PBF $\lambda$ = 1.5 | 0.017 | 0.011 | 0.010 | 0.010 | 0.018 | 0.012 | 0.011 | 0.008 |
| CW with PBF $\lambda$ = 2.0 | 0.020 | 0.013 | 0.012 | 0.012 | 0.02 | 0.013 | 0.012 | 0.008 |
| CW with PBF $\lambda$ = 4.0 | 0.033 | 0.022 | 0.017 | 0.015 | 0.029 | 0.020 | 0.016 | 0.009 |
| CW with PBF $\lambda$ = 6.0 | 0.046 | 0.030 | 0.024 | 0.019 | 0.038 | 0.028 | 0.020 | 0.011 |
| CW | 0.010 | 0.007 | 0.008 | 0.008 | 0.011 | 0.007 | 0.008 | 0.007 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.009 | 0.010 | 0.011 | 0.009 | 0.009 | 0.006 | 0.008 | 0.004 |
| CW with PBF $\lambda$ = 1.5 | 0.010 | 0.011 | 0.012 | 0.010 | 0.01 | 0.007 | 0.008 | 0.004 |
| CW with PBF $\lambda$ = 2.0 | 0.011 | 0.011 | 0.014 | 0.010 | 0.011 | 0.007 | 0.009 | 0.005 |
| CW with PBF $\lambda$ = 4.0 | 0.016 | 0.015 | 0.017 | 0.012 | 0.014 | 0.012 | 0.011 | 0.005 |
| CW with PBF $\lambda$ = 6.0 | 0.021 | 0.020 | 0.020 | 0.015 | 0.018 | 0.015 | 0.013 | 0.007 |
| CW | 0.008 | 0.008 | 0.009 | 0.009 | 0.007 | 0.005 | 0.007 | 0.004 |

*Notes*: 1. Table A.4 is equivalent to Table 1 in the main body of the paper with the only difference that the nominal size in Table A.4 is 1% and not 5% as in Table 1.

2. See notes to Table 1 for further details.

**Table A.5**

Size-Adjusted-Power: One-step-ahead forecasts, nominal size = 1%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.165 | 0.266 | 0.365 | 0.764 | 0.168 | 0.321 | 0.494 | 0.930 |
| CW with PBF $\lambda$ = 1.5 | 0.168 | 0.272 | 0.363 | 0.764 | 0.174 | 0.318 | 0.499 | 0.930 |
| CW with PBF $\lambda$ = 2.0 | 0.167 | 0.278 | 0.363 | 0.763 | 0.184 | 0.324 | 0.498 | 0.931 |
| CW with PBF $\lambda$ = 4.0 | 0.169 | 0.293 | 0.383 | 0.759 | 0.193 | 0.332 | 0.505 | 0.933 |
| CW with PBF $\lambda$ = 6.0 | 0.175 | 0.303 | 0.388 | 0.768 | 0.209 | 0.338 | 0.521 | 0.934 |
| CW | 0.155 | 0.261 | 0.361 | 0.766 | 0.156 | 0.321 | 0.477 | 0.929 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.055 | 0.091 | 0.147 | 0.306 | 0.058 | 0.110 | 0.202 | 0.592 |
| CW with PBF $\lambda$ = 1.5 | 0.055 | 0.090 | 0.146 | 0.307 | 0.061 | 0.111 | 0.202 | 0.593 |
| CW with PBF $\lambda$ = 2.0 | 0.055 | 0.088 | 0.145 | 0.306 | 0.061 | 0.114 | 0.201 | 0.595 |
| CW with PBF $\lambda$ = 4.0 | 0.056 | 0.090 | 0.147 | 0.309 | 0.062 | 0.117 | 0.206 | 0.599 |
| CW with PBF $\lambda$ = 6.0 | 0.062 | 0.093 | 0.146 | 0.310 | 0.064 | 0.121 | 0.213 | 0.598 |
| CW | 0.050 | 0.090 | 0.147 | 0.307 | 0.055 | 0.106 | 0.200 | 0.589 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.475 | 0.582 | 0.653 | 0.978 | 0.543 | 0.724 | 0.811 | 1.000 |
| CW with PBF $\lambda$ = 1.5 | 0.489 | 0.584 | 0.656 | 0.978 | 0.553 | 0.729 | 0.815 | 1.000 |
| CW with PBF $\lambda$ = 2.0 | 0.500 | 0.590 | 0.663 | 0.979 | 0.563 | 0.733 | 0.818 | 1.000 |
| CW with PBF $\lambda$ = 4.0 | 0.530 | 0.610 | 0.677 | 0.980 | 0.603 | 0.751 | 0.827 | 1.000 |
| CW with PBF $\lambda$ = 6.0 | 0.558 | 0.623 | 0.688 | 0.980 | 0.626 | 0.771 | 0.839 | 1.000 |
| CW | 0.441 | 0.565 | 0.634 | 0.977 | 0.507 | 0.707 | 0.798 | 1.000 |

*Notes*: 1. Table A.5 is equivalent to Table 2 in the main body of the paper with the only difference that the nominal size in Table A.5 is 1% and not 5% as in Table 2.
2. See notes to Table 1 for further details.

**Table A.6**

Raw Power: One-step-ahead forecasts, nominal size = 1%.

| | Panel A: Rolling regressions | | | | Panel B: Recursive regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | DGP 1 ($R$ = 120) | | | | DGP 1 ($R$ = 120) | | | |
| Test | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 | $P$ = 120 | $P$ = 240 | $P$ = 360 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.156 | 0.243 | 0.323 | 0.703 | 0.186 | 0.329 | 0.479 | 0.918 |
| CW with PBF $\lambda$ = 1.5 | 0.172 | 0.264 | 0.343 | 0.716 | 0.206 | 0.350 | 0.498 | 0.922 |
| CW with PBF $\lambda$ = 2.0 | 0.190 | 0.280 | 0.359 | 0.725 | 0.223 | 0.369 | 0.512 | 0.924 |
| CW with PBF $\lambda$ = 4.0 | 0.248 | 0.340 | 0.424 | 0.757 | 0.287 | 0.423 | 0.561 | 0.932 |
| CW with PBF $\lambda$ = 6.0 | 0.301 | 0.388 | 0.470 | 0.785 | 0.343 | 0.467 | 0.599 | 0.939 |
| CW | 0.119 | 0.198 | 0.283 | 0.676 | 0.147 | 0.285 | 0.437 | 0.911 |
| | DGP 2 ($R$ = 120) | | | | DGP 2 ($R$ = 120) | | | |
| | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 | $P$ = 85 | $P$ = 170 | $P$ = 340 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.070 | 0.089 | 0.141 | 0.301 | 0.080 | 0.111 | 0.198 | 0.555 |
| CW with PBF $\lambda$ = 1.5 | 0.078 | 0.099 | 0.148 | 0.310 | 0.090 | 0.122 | 0.211 | 0.560 |
| CW with PBF $\lambda$ = 2.0 | 0.087 | 0.108 | 0.159 | 0.320 | 0.098 | 0.131 | 0.220 | 0.566 |
| CW with PBF $\lambda$ = 4.0 | 0.127 | 0.143 | 0.194 | 0.357 | 0.141 | 0.168 | 0.253 | 0.600 |
| CW with PBF $\lambda$ = 6.0 | 0.160 | 0.172 | 0.229 | 0.386 | 0.168 | 0.204 | 0.234 | 0.610 |
| CW | 0.051 | 0.071 | 0.117 | 0.279 | 0.061 | 0.092 | 0.181 | 0.542 |
| | DGP 3 ($R$ = 240) | | | | DGP 3 ($R$ = 240) | | | |
| | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 | $P$ = 120 | $P$ = 180 | $P$ = 240 | $P$ = 1000 |
| CW with PBF $\lambda$ = 1.0 | 0.460 | 0.581 | 0.660 | 0.977 | 0.524 | 0.682 | 0.781 | 0.999 |
| CW with PBF $\lambda$ = 1.5 | 0.487 | 0.595 | 0.673 | 0.978 | 0.548 | 0.699 | 0.791 | 0.999 |
| CW with PBF $\lambda$ = 2.0 | 0.512 | 0.613 | 0.682 | 0.979 | 0.572 | 0.714 | 0.800 | 0.999 |
| CW with PBF $\lambda$ = 4.0 | 0.571 | 0.659 | 0.714 | 0.982 | 0.641 | 0.762 | 0.835 | 0.999 |
| CW with PBF $\lambda$ = 6.0 | 0.616 | 0.690 | 0.742 | 0.983 | 0.686 | 0.798 | 0.859 | 0.999 |
| CW | 0.399 | 0.531 | 0.627 | 0.975 | 0.459 | 0.627 | 0.748 | 0.999 |

*Notes*: 1. Table A.6 is equivalent to Table 3 in the main body of the paper with the only difference that the nominal size in Table A.6 is 1% and not 5% as in Table 3.

2. See notes to Table 1 for further details.

## Appendix B - An Intuition for More Power with a Simple i.i.d Example[9]

Here we show with a simple example in the context of i.i.d random variables why a simple t-statistic multiplied by our *"power-booster-factor"* may have higher size-adjusted-power. Imagine the following setup:

$$H_0 : \mu \leq 0 \text{ v/s } H_A : \mu > 0.$$

This is a one-sided-version of a traditional zero-mean test. Let us assume we have a sample of $T$ i.i.d. random variables $X_i$ all with the same expected value $\mu$ and the same variance $\sigma^2$. We would like to compare the behavior of the traditional test statistic

$$t_1 \equiv \sqrt{N}\left[\frac{\bar{X}}{\hat{\sigma}}\right].$$

With that of the same t-statistic multiplied by a version of our *"power-booster-factor"*:

$$t_2 \equiv \sqrt{T}\left[\frac{\bar{X}}{\hat{\sigma}}\right][1+\bar{X}]^{\lambda}.$$

Here $\bar{X}$ represents the sample average of $X_i$, $i = 1, \ldots T$ and $\hat{\sigma}$ is a consistent estimator of $\sigma$. Besides $\lambda \geq 1$.

By the central limit theorem we have that

$$\sqrt{T}\left[\bar{X} - \mu\right] \to N(0, \sigma^2).$$

Let us consider the differentiable function

$$g(\bar{X}) = \bar{X}\left[1+\bar{X}\right]^{\lambda}.$$

According to the Delta method we should have

$$\sqrt{T}\left[g(\bar{X}) - g(\mu)\right] \to N(0, \sigma^2[g'(\mu)]^2).$$

Which is equivalent to

$$\sqrt{T}\left[\bar{X}[1+\bar{X}]^{\lambda} - \mu[1+\mu]^{\lambda}\right] \to N\left(0, \sigma^2[[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1}]^2\right).$$

So, for large enough T we have

$$\sqrt{T}\left[\bar{X}[1+\bar{X}]^{\lambda}\right] \approx N\left(\sqrt{T}\mu[1+\mu]^{\lambda}, \sigma^2[[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1}]^2\right).$$

This means that for large enough $T, N\left(\sqrt{T}\mu[1+\mu]^{\lambda}], \sigma^2[[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1}]^2\right)$ is a good approximation for the distribution of $\sqrt{T}\left[\bar{X}[1+\bar{X}]^{\lambda}\right]$, which also indicates that

$$\sqrt{T}\left[\frac{\bar{X}[1+\bar{X}]^{\lambda}}{\hat{\sigma}}\right] \approx N\left(\frac{\sqrt{T}\mu[1+\mu]^{\lambda}}{\sigma}, [[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1}]^2\right).$$

So the distribution of $\sqrt{T}\left[\frac{\bar{X}[1+\bar{X}]^{\lambda}}{\hat{\sigma}}\right]$ is well approximated by $N\left(\frac{\sqrt{T}\mu[1+\mu]^{\lambda}}{\sigma}, [[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1}]^2\right)$.

[9] We are grateful to a reviewer for a comment which led us to develop this example.

In addition, we know that the traditional t-statistic $t_1$ is well approximated by $N\left(\frac{\sqrt{T}\mu}{\sigma}, 1\right)$.

Under the null, when $\mu = 0$, the distributions of both statistics, $t_1$ and $t_2$ are well approximated by standard normal distributions. But under the alternative, when $\mu > 0$ the distribution of $t_2$ is centered to the right of the distribution of $t_1$ because

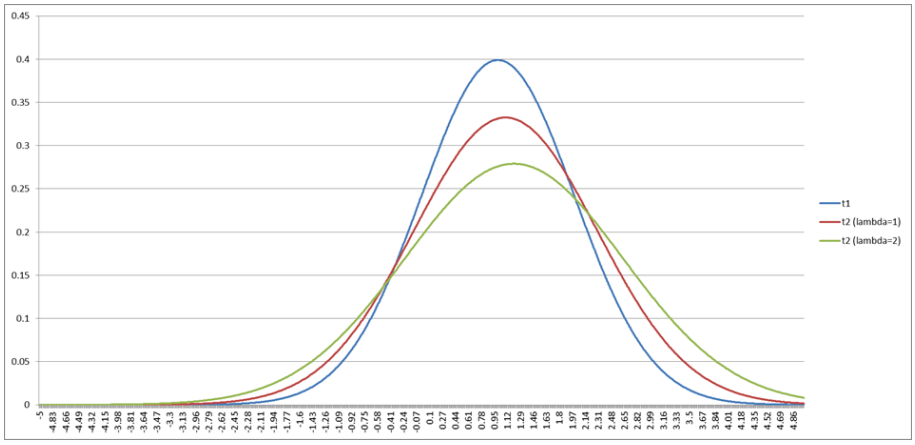
$$\frac{\sqrt{T}\mu[1+\mu]^{\lambda}]}{\sigma} > \frac{\sqrt{T}\mu}{\sigma}.$$

Furthermore, the approximated distribution of $t_2$ has higher variance relative to the approximated distribution of $t_1$ because

$$[1+\mu]^{\lambda} + \lambda\mu[1+\mu]^{\lambda-1} > 1.$$

Figure B.1 depicts the approximate distributions of $t_1$ and $t_2$. We picked two values of $\lambda$ for the graphical representation: $\lambda = 1$ and $\lambda = 2$. For the construction of Figure 1 we consider $T = 100$, $\sigma = 1$, $\mu = 0.1$. The vertical line in the graph shows the 5% critical value for an asymptotically normal one-sided-test (1.645). The area below the curves to the right of the vertical line represents the power of the different implicit tests. The usual t statistic $t_1$ has the lowest power in the picture: 25.9%. With the aid of the power booster factor and $\lambda = 1$ we have the red line with a power of 32.5%. Finally the green line represents the test with the *"power-booster-factor"* and $\lambda = 2$. This curve has the highest power in this scenario: 38.0%.

Interestingly, Figure B.1 looks pretty similar to Figure 2 in the main body of the paper, which depicts the CW test and the test with the *"power-booster-factor"* under the alternative. The Kernel distribution of CW has lower variance and it is centered to the left of the other distribution, which is the same implication that we get with the delta method in this simple i.i.d. case. Figure B.1 illustrates why the *"power-booster-factor"* could be useful to improve finite sample power of traditional one-sided tests.




*Notes*: $t_1$ is a standard t-statistic, while $t_2$ is the same t-statistic but multiplied by a version of our *"power-booster-factor"*.

**Figure B.1.** Approximate distributions of $t_1$ and $t_2$ under the alternative $\mu > 0$.

## References


Campbell, J. Y. (2001). Why long horizons? A study of power against persistent alternatives. *Journal of Empirical Finance 8(5)*, 459-491.

Ciccarelli, M., and Mojon, B. (2010). Global Inflation. *The Review of Economics and Statistics 92(3)*, 524-535.

Chen, Y., Rogoff, K. S., and Rossi, B. (2010). Can Exchange Rates Forecast Commodity Prices? *The Quarterly Journal of Economics 125(3)*, 1145-1194.

Clark, T., and McCracken, M. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105(1)*, 85-110.

Clark, T., and McCracken, M. (2005). Evaluating Direct Multistep Forecasts. *Econometric Reviews 24(4)*, 369-404.

Clark, T., and McCracken, M. (2013). Evaluating the Accuracy of Forecasts from Vector Autoregressions. In T. Fomby, L. Killian, and A. Murphy (Eds.), *Advances in Econometrics, vol. 32: VAR Models in Macroeconomics ? New Developments and Applications: Essays in Honor of Christopher A. Sims* (pp. 117-168). Bingley: Emerald Group Publishing.

Clark, T., and West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics 135(1-2)*, 155-186.

Clark, T., and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138(1)*, 291-311.

Diebold, F., and Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics 13(3)*, 253-263.

Duncan, R., and Martínez-García, E. (2015). Forecasting Local Inflation with Global Inflation: When Economic Theory Meets the Facts. Federal Reserve Bank of Dallas Working Paper No. 235.

Giacomini, R., and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica 74(6)*, 1545-1578.

Hakkio, C. (2009). Global Inflation Dynamics. Federal Reserve Bank of Kansas City Research Working Paper No. 09-01.

Hamilton, J. (1994). *Time Series Analysis*. New Jersey: Princeton University Press.

Inoue, A., and Kilian, L. (2005). In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews 23(4)*, 371-402.

Kabukcuoglu, A., and Martínez-García, E. (2018). Inflation as a global phenomenon—Some implications for inflation modeling and forecasting. *Journal of Economic Dynamics and Control 87*, 46-73.

Mankiw, N. G., and Shapiro, M. D. (1986). Do We Reject Too Often? Small Sample Properties of Tests of Rational Expectations Models. *Economics Letters 20(2)*, 139-145.

McCracken, M. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics 140(2)*, 719-752.

Medel, C., Pedersen, M., and Pincheira, P. (2016). The Elusive Predictive Ability of Global Inflation. *International Finance 19(2)*, 120-146.

Morales-Arias, L., and Moura, G. V. (2013). A conditionally heteroskedastic global inflation model. *Journal of Economic Studies 40(4)*, 572-596.

Nelson, C. R., and Kim, M. J. (1993). Predictable Stock Returns: The Role of Small Sample Bias. *Journal of Finance 48(2)*, 641-661.

Newey, W. K., and West, K. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55(3)*, 703-708.

Newey, W. K., and West, K. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *Review of Economic Studies 61(4)*, 631-653.

Pincheira, P. M. (2013). Shrinkage Based Tests of Predictability. *Journal of Forecasting 32(4)*, 307-332.

Pincheira, P., and Gatty, A. (2016). Forecasting Chilean Inflation with International Factors. *Empirical Economics 51(3)*, 981-1010.

Pincheira, P. M., and West, K. D. (2016). A comparison of some out-of-sample tests of predictability in iterated multi-step-ahead forecasts. *Research in Economics 70(2)*, 304-319.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54(3)*, 375-421.

Tauchen, G. (2001). The bias of tests for a risk premium in forward exchange rates. *Journal of Empirical Finance 8(5)*, 695-704.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica 64(5)*, 1067-1084.

West, K. D. (2006). Forecast Evaluation. In G. Elliott, C. Granger and A. Timmerman (Eds.), *Handbook of Economic Forecasting, Vol. 1* (pp. 100-134). Amsterdam: Elsevier.

White, H. (2001). *Asymptotic Theory for Econometricians: Revised Edition.* San Diego and London: Academic Press.