

A Predictability Test for a Small Number of Nested Models

Eleonora Granziera*, Kirstin Hubrich**, Hyungsik Roger Moon†

September 29, 2011

Abstract

In this paper we introduce tests of Likelihood Ratio types for one sided multivariate hypothesis to evaluate the null that a parsimonious model performs equally well as a small number of models which nest the benchmark. We show that the limiting distributions of the test statistics are non standard. For critical values we consider two approaches: (i) bootstrapping and (ii) simulations assuming normality of the mean square prediction error (MSPE) difference. The size and the power performance of the tests are compared via Monte Carlo experiments with two existing tests proposed in Hubrich and West (2010): a chi-squared test and the maximum of t-statistic test. We find that all tests are well sized for one step ahead forecasts; for multi-step forecasts the normal approximation delivers grossly oversized tests, while the bootstrap provides with smaller size distortions. The experiments on the power reveal that the chi-squared test performs last while the ranking between the likelihood-ratio type test and the max-t stat depends on the simulation settings. Last, we apply our test to draw conclusions about the predictive ability of a Phillips type curve for the US core inflation.

Keywords: Out-of sample, point-forecast evaluation, multi-model comparison, predictive ability, direct multi-step forecasts, fixed regressors bootstrap.

We thank Raffaella Giacomini, Peter Hansen, Søren Johansen, Michael McCracken, Hashem Pesaran, Norman Swanson, Kenneth West, participants at USC, BoC, Conference in Honor of Hal White, NASM 2011, AMES 2011, ESEM 2011 and EUI Conference for helpful comments and suggestions.

* Bank of Canada, egranziera@bankofcanada.ca; ** European Central Bank, Research Department, kirstin.hubrich@ecb.europa.eu; † University of Maryland and USC, moonr@usc.edu

1 Introduction

Evaluation of forecast accuracy usually requires to compare the expected loss of the forecasts obtained from a set of models of interest. Testing whether the models provide the same forecast performance represents a test of equal predictive ability.

Early literature focuses on comparing non-nested models. Diebold and Mariano (1995) and West (1996) suggest a framework to test for equal predictive ability in the case of pairwise model comparison of nested models. White (2000) suggests a test for superior predictive ability for a large number of models in a non-nested framework. Corradi and Swanson (2007) modify the framework in White (2000) to allow for parameter errors to enter the asymptotic distribution; again the benchmark model should be nonnested in at least one of the competing models.

However, in many applications the benchmark might be a parsimonious model obtained by imposing zero restrictions on the coefficients associated with the predictors in the alternative models. Examples include: Cooper and Gulen (2004), Guo (2006), Goyal and Welch (2008) for stock market predictability, Stock and Watson (1999), Hubrich (2005), Hendry and Hubrich (2011) for inflation, Stock and Watson (2003), Ravazzolo and Rothman (2010), Andersson et al. (2011) for GDP growth. In this case, it is well known in the literature that many equal predictive ability tests developed for non-nested models cannot be used due to failure of the rank condition¹ (*e.g.*, See West (2006)).

In more recent years the analysis of pairwise model comparison of nested models has been the object of many studies. Chao et al. (2001) derive out-of-sample Granger's causality tests² that have standard normal limiting distributions for one-step ahead predictions. For multi-step forecasts obtained with the direct method and nonlinear least squares parameter estimation formal characterization of the limiting distributions has been attained by Clark and McCracken (2001, 2005) and McCracken (2007). In this environment the test statistic to evaluate the null of equal predictive ability is derived as functionals of Brownian motions and is asymptotically pivotal under certain additional conditions. Clark and West (2006, 2007), thereafter CW, argue that for nested models the finite sample mean square prediction error (MSPE) difference is negative and they introduce an adjustment term to center the statistic around zero to get correctly sized tests. They also provide Monte Carlo evidence supporting their suggested procedure.

¹Under the null of equal predictive accuracy the errors of the different models are the same and therefore the covariance matrix of the estimator is not full rank.

²Out-of-sample Granger causality tests can be interpreted as predictability tests for nested models.

In this paper we extend the nested pairwise model comparison set-up to a nested multiple model comparison. The main objective of the paper is to test out-of-sample equal predictive ability with multiple models when a benchmark model is nested by the small number of remaining models.³ In the existing literature Hubrich and West (2010) (hereafter HW) consider this setup⁴ and propose two approaches: one is to directly extend the pairwise model comparison in Diebold and Mariano (1995) and West (1996), (DMW) to a chi-squared statistic and the other is to take the maximum of t-statistics of all the pairwise MSPE differences, resulting in inference based on the maximum correlated normals. Both tests are Wald-type tests and they adjust the MSPE differences as advocated in CW for pairwise model forecast comparison.

The main contribution of the paper is to propose an alternative test to the ones in HW. When the null model is nested by the alternative models, we first notice that the MSPE differences are zeros under the null of equal predictability while they are non-negative under the alternative. Then, by treating the MSPE differences as a multivariate parameter to test, the problem of testing for equal predictability translates into testing a multivariate parameter that takes one-sided values. In this paper we formulate the testing problem as just described and we propose one-sided likelihood ratio (hereafter LR) type predictability tests for the comparison of a small number of models nesting the benchmark model. The LR test statistic depends on the structure of the alternative models. We distinguish among three different cases according to the structure of the alternative models considered: (i) when the models are nested within each other, (ii) when there is no nesting relation among the alternative models, and (iii) when the models can be grouped such that within each group the models are nested, but there is no nesting relation among groups. We derive the asymptotic distribution of the tests and find that they depend on characteristics of the predictors. This implies that one needs to tabulate the critical values for every application to use the asymptotic distribution for testing. As an alternative to avoid this problem we consider two approaches, (i) bootstrapping and (ii) simulations based on normal approximation of the MSPE difference estimates.

The finite sample size and power properties of the tests are evaluated via extensive Monte Carlo simulations for one and four-step ahead forecasts. The tests we compare include three

³A rigorous definition of ‘small’ is not provided, but as a practical rule we suggest that the number of models should be smaller than the size of the out-of sample period.

⁴Tests for multiple forecast encompassing are provided in Harvey and Newbold (2000) and in Clark and McCracken (2011). The tests suggested in Hubrich and West (2010) are also equivalent to testing forecast encompassing for multiple models simultaneously.

LR type tests, the max-t test of HW, and the Wald type test. Our Monte Carlo investigation reveals that, as previously found by Hubrich and West, the use of simulated critical values based on normal approximation of the MSPE differences performs well overall in terms of the size in the case of one-step ahead forecasts. However, for four step ahead forecasts and for small out-of-sample the tests are grossly oversized, although the size distortions decrease as size of the out of sample increases. For critical values derived through the fixed regressors bootstrap, as in Clark and McCracken (2011), the tests are correctly sized for one step ahead forecasts and the size distortions for the longer forecast horizons are not severe. As far as the power of the test is concerned, regardless which critical values we use, the chi-square test performs poorly in terms of power as it disregards the one-sided nature of the test. The ranking between the likelihood-ratio type test and the max t-statistics test depends on the simulation settings. This result is expected given that there is no uniformly most powerful test for multivariate one sided hypothesis about linear equality constraints. Relying on the bootstrap rather than on the assumption of normality to compute the critical values can decrease the power of the tests, especially in small samples.

As an illustrative application, we evaluate equal predictive ability for forecasting the US CPI core yearly inflation rate among an AR(1) model as a benchmark and three other alternative models that extend the AR(1) model by including extra predictors. Evidence against the null of equal predictive ability is mixed and it varies not only across samples, but it also depends on the test considered and on the method used to obtain the critical values. Then, the simulation results provide us with some guidance on the most appropriate tests and critical values to consider in order to draw conclusions about the predictive ability of a Phillips type curve for US core inflation.

The outline of the paper is as follows: section 2 introduces the notation and the forecasting environment and presents the tests. Section 3 provides procedures for inference based on the tests. In section 4 the Monte Carlo simulation experiments are described and the size and power properties of the tests are discussed. An empirical application of the test, forecasting core US inflation, is presented in Section 5. Section 6 concludes.

2 Basic Framework

Suppose that $\{y_{s+\tau}, x_s\}_{s=1}^t$ are observed stationary time series variables at each forecast origin $t = T, \dots, T + P - \tau$ and that $x_{m,t}$ are the predictors that belong to x_t , for $m = 0, 1, \dots, M$. Notation m is used to denote a forecasting model. The benchmark model is denoted by

$m = 0$ and the alternative models by $m = 1, \dots, M$.

Suppose that one is interested in forecasting a scalar $y_{t+\tau}$, $\tau \geq 1$, using $M + 1$ linear⁵ models:

$$\begin{aligned} y_{t+\tau} &= x'_{0,t}\beta_0^0 + u_{0,t+\tau} \\ &\vdots \\ y_{t+\tau} &= x'_{m,t}\beta_m^0 + u_{m,t+\tau} \\ &\vdots \\ y_{t+\tau} &= x'_{M,t}\beta_M^0 + u_{M,t+\tau}, \end{aligned} \tag{1}$$

where $x'_{m,t}\beta_m^0$ is the linear projection of $y_{t+\tau}$ on the predictor $x_{m,t}$ and $u_{m,t+\tau}$ denotes the forecast error satisfying $E(u_{m,t+\tau}x_{m,t}) = 0$ for $m = 0, 1, \dots, M$. Notice that the time series of the linear projection errors $u_{m,t+\tau}$ could be serially correlated, in particular for multistep forecasts. For $\tau \geq 2$, we allow the forecast errors to follow a $MA(\tau - 1)$ process. We assume the parameters β_m to be constant over time and we do not allow for structural breaks.

Denote by $\hat{y}_{0,t+\tau}, \dots, \hat{y}_{m,t+\tau}, \dots, \hat{y}_{M,t+\tau}$ the τ -period ahead forecasts obtained from the estimated models either through the expanding window or the rolling scheme⁶ for $t = T, \dots, T + P - \tau$. Here the total sample size is $T + P$, T is the size of the sample used to generate the initial estimates, P is the number of the observations used for out-of-sample evaluation. We consider the case where the benchmark model is nested by every alternative model by imposing the restriction that $x_{0,t}, \dots, x_{M,t}$ are vectors of predictors such that $x_{0,t} = x_{01,t}$ is of dimension $k_0 \times 1$ and $x_{m,t} = (x'_{01,t}, x'_{m2,t})'$ is of dimension $k_m \times 1$ with $k_0 < k_m$.

The main goal of the paper is to test for the null hypothesis that the parsimonious model, model 0, performs equally well as a larger model, say model m , $m \in \{1, \dots, M\}$. Under the null hypothesis, model 0 is the true model and hence each model m includes $k_m - k_0$ excess parameters:

$$\beta_m^0 = (\beta_0^{0r}, \mathbf{0}'_{k_m - k_0})',$$

for all $m = 1, \dots, M$. Moreover, under the null hypothesis the errors are identical $u_{0,t+\tau} = u_{1,t+\tau} = \dots = u_{M,t+\tau}$. Under the alternative, however, the additional parameters estimated

⁵Refer to Corradi and Swanson (2002) for an out of sample predictive accuracy test where the alternative model is unknown and (non)linear.

⁶In the expanding window scheme the size of the estimation sample grows while in the rolling scheme it stays constant.

are non-zero in population.

Following West (2006) and CW for pairwise comparisons and HW for multiple model comparisons we denote as $f_{m,t+\tau}$ the difference of the loss functions between the benchmark and alternative model m . In this paper, we consider $f_{m,t+\tau}$ to be the difference in the squared prediction errors (SPE): $f_{m,t+\tau} = u_{0,t+\tau}^2 - u_{m,t+\tau}^2$, where $u_{m,t+\tau} = y_{t+\tau} - y_{m,t+\tau}^f$ and $y_{m,t+\tau}^f$ is the τ -step ahead forecast from model m when the parameters of the models are set to their population values. Collect the SPE differences in the vector $f_{t+\tau}$:

$$f_{t+\tau} = (f_{1,t+\tau}, \dots, f_{M,t+\tau})'$$

Define μ to be the expected value of the difference in the SPE, i.e. the expected value of $f_{t+\tau}$:

$$\mu = E(f_{t+\tau}) = (\sigma_0^2 - \sigma_1^2, \dots, \sigma_0^2 - \sigma_M^2)'$$

with $\sigma_m^2 \equiv E(u_{m,t+\tau}^2)$ being the population variance of the forecast error, which is assumed to be a stationary process.

Let $\hat{u}_{m,t+\tau} = y_{t+\tau} - \hat{y}_{m,t+\tau}$ be the τ -step ahead forecast error from the estimated model m . The sample analogs of $f_{m,t+\tau}$ and $f_{t+\tau}$, denoted by $\hat{f}_{m,t+\tau}$ and $\hat{f}_{t+\tau}$, are given respectively by:

$$\hat{f}_{m,t+\tau} = (y_{t+\tau} - \hat{y}_{0,t+\tau})^2 - (y_{t+\tau} - \hat{y}_{m,t+\tau})^2 = (\hat{u}_{0,t+\tau})^2 - (\hat{u}_{m,t+\tau})^2$$

and

$$\hat{f}_{t+\tau} = (\hat{f}_{1,t+\tau}, \dots, \hat{f}_{M,t+\tau})'$$

The sample counterpart of μ , the sample mean SPE (MSPE), is given by:

$$\bar{f} = (P - \tau + 1)^{-1} \left(\sum_{t=T}^{T+P-\tau} \hat{f}_{1,t+\tau}, \dots, \sum_{t=T}^{T+P-\tau} \hat{f}_{m,t+\tau}, \dots, \sum_{t=T}^{T+P-\tau} \hat{f}_{M,t+\tau} \right)'$$

According to CW, under the null that model 0 is the correctly specified model the sample MSPE from the parsimonious model will generally be lower than the sample MSPE from the alternative model, so it may be the case that $(P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} (\hat{f}_{m,t+\tau}) < 0$ in finite samples. To improve the finite sample properties, they suggest to use the adjusted sample MSPE to center it around zero, as

$$\hat{f}_{m,t+\tau}^{adj} = \hat{f}_{m,t+\tau} + (\hat{y}_{0,t+\tau} - \hat{y}_{m,t+\tau})^2,$$

where $(\hat{y}_{0,t+\tau} - \hat{y}_{m,t+\tau})^2$ is the adjustment term. A simple algebra shows that:

$$\begin{aligned}\hat{f}_{m,t+\tau}^{adj} &= \hat{u}_{0,t+\tau}^2 - \hat{u}_{m,t+\tau}^2 + (\hat{y}_{0,t+\tau} - \hat{y}_{m,t+\tau})^2 \\ &= \hat{u}_{0,t+\tau}^2 - \hat{u}_{m,t+\tau}^2 + (\hat{u}_{m,t+\tau} - \hat{u}_{0,t+\tau})^2 \\ &= 2\hat{u}_{0,t+\tau}(\hat{u}_{0,t+\tau} - \hat{u}_{m,t+\tau}).\end{aligned}\tag{2}$$

Then, (2) establishes the equivalence between adjusted MSPEs suggested in Clark and West (2007) and pairwise model encompassing test statistics as in Harvey et al. (1998) or Clark and McCracken (2001).⁷ This implies that HW's proposed test statistics are encompassing tests for small nested model sets, and a similar interpretation can be given to our LRT tests. Analogous quantities defined above for $\hat{f}_{m,t+\tau}$ can be derived from $\hat{f}_{m,t+\tau}^{adj}$:

$$\hat{f}_{t+\tau}^{adj} = \left(\hat{f}_{1,t+\tau}^{adj}, \dots, \hat{f}_{M,t+\tau}^{adj} \right)' \text{ and } \bar{f}^{adj} = (P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} \hat{f}_{t+\tau}^{adj}.$$

Following Hubrich and West's suggestion for multiple model comparison, we define

$$\mu_m^{adj} = \mu_m + E \left(y_{0,t+\tau}^f - y_{m,t+\tau}^f \right)^2.$$

In the Not-for-Publication Appendix we show that

$$\mu_m^{adj} = 2\mu_m,$$

so we conclude that in population the adjustment does not alter the nature of the problem stated by the unadjusted MSPE. We will specify the null hypothesis⁸ as $H_0 : \mu = 0$, or equivalently $H_0 : \mu^{adj} = 0$, while the specification of the alternative hypothesis will depend on the assumptions about the nesting structure of the alternative models. In the outline of our test we will distinguish three cases: (i) when the models are nested within each other, (ii) when there is no nesting relation between the alternative models, and (iii) when the models are nested within groups. The structure of the alternative models will determine the structure of the alternative hypothesis and therefore it will affect the power properties of the test.

⁷Note that Harvey et al. (1998) do abstract from parameter uncertainty so seems oriented towards a different class of applications.

⁸Giacomini and White (2006) suggest tests of conditional predictive ability where the null is expressed in terms of sample moments rather than population moments.

2.1 Case (i): Alternative Models Nested within Each Other

We characterize the case in which each model $m - 1$ is nested in model m by imposing that model m includes $k_m - k_{m-1}$ additional regressors: $x_{m,t} = (x'_{m-1,t}, X'_{m,t})$ so that $k_0 < \dots < k_m < \dots < k_M$.

Given the structure of the problem, if model m^* is the true model, then for models $m = 1, \dots, m^* - 1$, it will hold that $\sigma_{m^*}^2 < \sigma_{m^*-1}^2 < \dots < \sigma_1^2$ and hence $0 < \mu_1 = \sigma_0^2 - \sigma_1^2 < \dots < \sigma_0^2 - \sigma_{m^*-1}^2 = \mu_{m^*-1} < \sigma_0^2 - \sigma_{m^*}^2 = \mu_{m^*}$, while for models $m^* + 1, \dots, M$, it will be the case that $\sigma_{m^*}^2 = \sigma_{m^*+1}^2 = \dots = \sigma_M^2$, which implies $\mu_{m^*} = \mu_{m^*+1} = \dots = \mu_M$. Note this holds only for the case in which the set of regressors is progressively expanding with the models,⁹ meaning that for every model m the set of regressors in model $m - 1$ is a subset of the regressors in model m . Then the null and the alternative hypotheses can be expressed as:¹⁰

$$\begin{aligned} H_0 &: \mu = 0. \\ H_1 &: 0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_M, \mu \neq 0. \end{aligned} \tag{3}$$

Hence we test equal forecast accuracy versus the alternative that at least one of the models performs better than the benchmark. We consider a one-sided alternative as first suggested by Ashley et al. (1980) and subsequently assumed in many studies (CW, HW).

The test we propose to evaluate the null of equal predictive ability is a likelihood-ratio type test of the form:

$$\mathcal{T}_{LRT_D} = (P - \tau + 1) \bar{f}^{adj} \hat{V}^{-1} \bar{f}^{adj} - \min_{D\mu \geq 0} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu)$$

where

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ & & \ddots & \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

⁹A notable example is provided in the seminal paper by Meese and Rogoff (1983) on predictability of exchange rates.

¹⁰Equivalently, since this ordering is invariant to the introduction of the CW adjustment the null and alternative can be expressed with respect to μ^{adj} : $H_0 : \mu^{adj} = 0$ vs $H_1 : 0 \leq \mu_1^{adj} \leq \mu_2^{adj} \leq \dots \leq \mu_M^{adj}$, $\mu^{adj} \neq 0$.

is the first difference operator and

$$\hat{V} = \sum_{j=-\tau+1}^{\tau-1} \left[\frac{1}{P-\tau+1} \sum_{t:T \leq t, t+j \leq T+P-\tau} \left(\hat{f}_{t+\tau}^{adj} - \bar{f}^{adj} \right) \left(\hat{f}_{t+\tau+j}^{adj} - \bar{f}^{adj} \right)' \right] \quad (4)$$

is an estimator of the long-run variance of \bar{f}^{adj} , considering that the forecast errors have $MA(\tau-1)$ type serial dependence structures. Then, the alternative hypothesis in (3) can be expressed as

$$H_1 : D\mu \geq 0.$$

2.2 Case (ii): Non-Nested Alternative Models

In this case there is no nesting relation between the alternative models, but still each of them nests the benchmark. We test for:

$$H_0 : \mu = 0$$

against

$$H_1 : \mu_1 \geq 0, \dots, \text{ and } \mu_M \geq 0, \mu \neq 0. \quad (5)$$

Then, the associated LR type can be formulated as:

$$\mathcal{T}_{LRT_I} = (P - \tau + 1) \bar{f}^{adj}' \hat{V}^{-1} \bar{f}^{adj} - \min_{I\mu \geq 0} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu),$$

where \hat{V} is defined as in (4) and I is the $M \times M$ identity matrix.¹¹

2.3 Case (iii): Alternative Models Nested within Groups

Now we consider a general case. Suppose that the alternative models can be grouped according to the following relations: within each group the models are nested; however across different groups, the models are not nested. In particular, consider K groups such that within each group k : $\mu_{k,1} \leq \mu_{k,2} \leq \dots \leq \mu_{k,M_k}$, with M_k the number of models included in group k . Then, define \mathcal{A}_k by

$$\mathcal{A}_k = \left\{ \mu : 0 \leq \mu_{k,1} \leq \mu_{k,2} \leq \dots \leq \mu_{k,M_k} \right\}.$$

¹¹Note that the difference between the \mathcal{T}_{LRT_I} and \mathcal{T}_{LRT_D} statistics is the use of the I matrix instead of the difference operator matrix D to characterize the likelihood under the alternative.

For this case we propose a likelihood-ratio type test that combines the two tests outlined in the previous sections.

$$\mathcal{T}_{LRT} = \max_{1 \leq k \leq K} \left\{ (P - \tau + 1) \bar{f}^{adj'} \hat{V}^{-1} \bar{f}^{adj} - \min_{\mu \in \mathcal{A}_1} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu), \dots \right. \\ \left. \dots, (P - \tau + 1) \bar{f}^{adj'} \hat{V}^{-1} \bar{f}^{adj} - \min_{\mu \in \mathcal{A}_K} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu) \right\}$$

where

$$\mathcal{A}_k = \{ \mu \in \mathbb{R}^M : \mathcal{D}_k \mu \geq 0 \} \text{ for } k = 1, \dots, K,$$

$$\mathcal{D}_k = \begin{bmatrix} & & 0_{((\sum_{j=1}^k M_j) \times M_k)} & & \\ & & \text{---} & & \\ 0_{(M \times (\sum_{j=1}^k M_j))} & | & D_k & | & 0_{(M \times (\sum_{j=k+1}^K M_j))} \\ & & \text{---} & & \\ & & 0_{((\sum_{j=k+1}^K M_j) \times M_k)} & & \end{bmatrix},$$

and

$$D_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

2.4 Alternative Tests

We consider two alternative forecast accuracy tests for nested multi-model comparison considered in the existing literature both proposed by HW: the first one is a chi-squared test, the second one is the max t-statistic test.

HW considers a Wald-type test involving the statistic

$$\mathcal{T}_{chi} = (P - \tau + 1) \bar{f}^{adj'} \hat{V}^{-1} \bar{f}^{adj}$$

As discussed above, it uses adjusted MSPE as suggested by CW for pairwise model comparison in order to center the statistic around zero. This test does not take into account the one-sided nature of the alternative and therefore is expected to have low power. Hence, HW

exploits the one-sided nature of the alternative and suggest another test statistic:

$$\mathcal{T}_{\max t} = \max_{1 \leq m \leq M} \left\{ \sqrt{(P - \tau + 1)} \frac{\bar{f}_1^{adj}}{\sqrt{\hat{v}_1}}, \dots, \sqrt{(P - \tau + 1)} \frac{\bar{f}_M^{adj}}{\sqrt{\hat{v}_M}} \right\},$$

which is the maximum of the t-statistics where \hat{v}_m is the m^{th} diagonal element of \hat{V} .

3 Computation of Critical Values

In this section we discuss how to compute critical values for the test statistics of the previous section. For this, we first derive the limiting distribution of the test statistics. We show that the limit is not only a complicated functional of Brownian motion but also non-pivotal, which makes it difficult to use the critical values of the limit distribution. As alternatives, we consider two different approaches. The first method is the bootstrapping approach. The second method is to use the Gaussian approximation. Due to space limitation, we consider only the \mathcal{T}_{LRT} statistic since it covers a general case. It is straightforward to modify these approaches for the test statistics \mathcal{T}_{LRT_D} and \mathcal{T}_{LRT_I} .

3.1 Limiting Distribution

We derive the asymptotic distribution of the general test statistic

$$\begin{aligned} \mathcal{T}_{LRT} = \max_{1 \leq k \leq K} & \left\{ (P - \tau + 1) \bar{f}^{adj'} \hat{V}^{-1} \bar{f}^{adj} - \min_{\mu \in \mathcal{A}_1} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu), \dots \right. \\ & \left. \dots, (P - \tau + 1) \bar{f}^{adj'} \hat{V}^{-1} \bar{f}^{adj} - \min_{\mu \in \mathcal{A}_K} (P - \tau + 1) (\bar{f}^{adj} - \mu)' \hat{V}^{-1} (\bar{f}^{adj} - \mu) \right\}. \end{aligned}$$

Then, we discuss how to compute the critical values based on the limiting distribution.

For this, we let x_t denote the k_x -vector of all the predictors that do not overlap. Denote J_m to be the $(k_m \times k_x)$ selection matrix such that $x_{m,t} = J_m x_t$ for $m = 0, 1, \dots, M$. Recall that under the null hypothesis, $u_{0,t} = u_{1,t} = \dots = u_{M,t}$. We let $u_t = u_{m,t}$ for all t . Let $h_{t+\tau} = x_t u_{t+\tau}$, $H_t = \frac{1}{t} \sum_{s=1}^{t-\tau} h_{t+\tau}$. Denote $\Sigma_x = E(x_t x_t')$ and $\Omega_h = \lim_T \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T E(h_t h_s')$. Define $\tilde{h}_t = \Omega_h^{1/2} h_t$. Denote $Q_m = \Omega_h^{1/2} (J_m' (J_m \Sigma_x J_m')^{-1} J_m - J_0' (J_0 \Sigma_x J_0')^{-1} J_0) \Omega_h^{1/2}$. We use $W(r)$ to denote the k_x dimensional Wiener process. We assume the following assumptions which are quite standard in the literature (e.g., Clark and McCracken (2001, 2005, 2011)),

McCracken (2007)).¹²

Assumption 1 *The coefficient β_m of Model m is estimated recursively by OLS:*

$$\hat{\beta}_{m,t} = \arg \min_{\beta_m} \frac{1}{t} \sum_{s=1}^{t-\tau} (y_{t+\tau} - \beta_m' x_{m,t})^2, \text{ for } m = 0, 1, \dots, M.$$

Denote $U_{t+\tau} = (h'_{t+\tau}, \text{vech}(x_t x_t' - E(x_t x_t')))'$.

Assumption 2 *(a) U_t is strictly stationary with $E(U_t) = 0$ and $E\|U_t\|^r < \infty$, for some $r > 8$. (b) $E(h_t h_{t-j}') = 0$ for all $j \geq \tau$. (c) $E(x_t x_t') > 0$. (d) For some $r > d > 2$, $\{U_t\}$ is strong mixing with mixing coefficients of size $-\frac{rd}{r-d}$. (e) The long run variance of U_t , $\lim_T \frac{1}{T} E \left[\left(\sum_{s=1}^{T-\tau} U_{s+\tau} \right) \left(\sum_{s=1}^{T-\tau} U_{s+\tau} \right)' \right]$ is a finite and positive definite matrix..*

Define the $(M \times M)$ matrix B such that its $(m, n)^{\text{th}}$ element is $\text{tr}(Q_m Q_n)$, that is,¹³

$$B = [\text{tr}(Q_m Q_n)]_{(m,n)}.$$

Assumption 3 *Assume that $\text{rank}(B) = M$.*

Assumption 4 *Assume that $\lim \frac{P}{T} = \lambda \in (0, \infty)$.*

Suppose that a is an M -vector and A is a (strictly) positive definite $M \times M$ matrix. Define the functional

$$\mu_k(a, A) = \arg \min_{\mu \in \mathcal{A}_k} (a - \mu)' A^{-1} (a - \mu).$$

Then, it is well known that

$$a' A^{-1} a - \min_{\mu \in \mathcal{A}_k} (a - \mu)' A^{-1} (a - \mu) = \mu_k(a, A)' A^{-1} \mu_k(a, A)$$

and $\mu_k(a, V)$ is continuous in (a, A) . (e.g., see page 213 of Silvapulle and Sen).

Define

$$\mathcal{F} = (\mathcal{F}_1, \dots, \mathcal{F}_M)', \quad \mathcal{F}_m = \int_1^{1+\lambda} W(r)' Q_m dW(r),$$

¹²Refer to these studies for a detailed discussion of the assumptions. Assumption 3 is specific of our environment and it ensures invertibility of \hat{V} .

¹³Note that in a special case where $\Omega_h = \Sigma_x = I_{k_x}$, $\text{tr}(Q_m Q_m)$ is $k_m - k_0$ and $\text{tr}(Q_m Q_n) = (\# \text{ of the common regressors in models } m \text{ and } n) - k_0$.

and

$$\mathcal{V} = \begin{bmatrix} \mathcal{V}_{11} & \cdots & \mathcal{V}_{1M} \\ \vdots & \ddots & \vdots \\ \mathcal{V}_{M1} & \cdots & \mathcal{V}_{MM} \end{bmatrix},$$

where

$$\mathcal{V}_{mn} = \int_1^{1+\lambda} W(r)' Q_m Q_n W(r) dr.$$

Theorem 1 *Assume Assumptions 1 – 4. Then,*

$$\mathcal{T}_{LRT} \Rightarrow \max_{1 \leq k \leq K} \left\{ \mu_1(\mathcal{F}, \mathcal{V})' \mathcal{V}^{-1} \mu_1(\mathcal{F}, \mathcal{V}), \dots, \mu_K(\mathcal{F}, \mathcal{V})' \mathcal{V}^{-1} \mu_K(\mathcal{F}, \mathcal{V}) \right\}.$$

Notice that the limiting distribution of \mathcal{T}_{LRT} is a functional of Brownian motion and it depends on the characteristics of the data generating process such as the out-of-sample to in-sample ratio and on the covariance matrix of the regressors that do not overlap. In order to use the critical values of the limit distribution of \mathcal{T}_{LRT} , one has to tabulate the critical values for every application using simulation and estimation. This approach may be burdensome and not practical. In what follows we consider two alternative approaches.

3.2 Bootstrap Approach

An alternative method to the asymptotic approach is the bootstrap method. In this paper, we consider the bootstrap procedure proposed by Clark and McCracken (2011) which is a variant of the wild fixed regressor bootstrap developed in Goncalves and Kilian (2004). A detailed procedure is:

Step 1: Compute $\hat{\beta} = \left(\sum_{s=1}^T x_s x_s' \right)^{-1} \left(\sum_{s=1}^T x_s y_{s+\tau} \right)$, the OLS estimator that uses the whole set of predictors x_t . Then, compute the residuals $\hat{u}_{s+\tau} = y_{s+\tau} - \hat{\beta}' x_{s+\tau}$, for $s = 1, \dots, T + P - \tau$.

Step 2: Fit $\hat{u}_{s+\tau}$ on an $MA(\tau - 1)$ process: $\hat{u}_{s+\tau} = \hat{\varepsilon}_{s+\tau} + \hat{\theta}_1 \hat{\varepsilon}_{s+\tau-1} + \cdots + \hat{\theta}_{\tau-1} \hat{\varepsilon}_{s+1}$. Simulate a sequence of *iid* $N(0, 1)$ random variables, $\eta_{s+\tau}$, where $s = 1, \dots, T + P - \tau$. Then, compute $\hat{u}_{s+\tau}^* = \eta_{s+\tau} \hat{\varepsilon}_{s+\tau} + \hat{\theta}_1 \eta_{s+\tau-1} \hat{\varepsilon}_{s+\tau-1} + \cdots + \hat{\theta}_{\tau-1} \eta_{s+1} \hat{\varepsilon}_{s+1}$, for $s = 1, \dots, T + P - \tau$.

Step 3: Estimate the null model by OLS: $\hat{\beta}_0 = \left(\sum_{s=1}^T x_{0,s} x_{0,s}' \right)^{-1} \left(\sum_{s=1}^T x_{0,s} y_{s+\tau} \right)$. Then, generate samples

$$y_{s+\tau}^* = x_{0,s}' \hat{\beta}_0 + \hat{u}_{s+\tau}^*$$

for $s = 1, \dots, T + P - \tau$.

Step 4: Using $\{y_{s+\tau}^*, x_s\}_{s=1, \dots, T+P-\tau}$, construct the test statistic \mathcal{T}_{LRT}^* .

Step 5: Repeat Steps 1–4 B times to compute $\mathcal{T}_{LRT}^{*(b)}$, $b = 1, \dots, B$. Compute the $(1 - \alpha)^{th}$ quantile of the empirical distribution of $\{\mathcal{T}_{LRT}^{*(b)}\}_b$ as the size α critical value.

Assumption 5 (a) Under the null, the forecast error u_t is an invertible MA $(\tau - 1)$ process generated by $u_t = \varepsilon_t + \theta_1^0 \varepsilon_{t-1} + \dots + \theta_{\tau-1}^0 \varepsilon_{t-\tau+1}$, where $\varepsilon_t \sim iid$ with $E(\varepsilon_t) = 0$, $E\|\varepsilon_t\|^r < \infty$, for some $r > 8$, and $\varepsilon_0 = \dots = \varepsilon_{1-\tau} = 0$. (b) Denote $\Theta(L; \theta) = 1 + \theta_1 L + \dots + \theta_{\tau-1} L^{\tau-1}$. Denote $\varepsilon_t(\theta, \beta) = \Theta(L; \theta)^{-1} u_t(\beta)$ with $u_0(\beta) = 0$, where $u_t(\beta) = y_{t+\tau} - \beta' x_t$. We assume that there exists an open neighborhood N of the true parameter (θ^0, β^0) and $r > 8$ such that $\sup_t \sup_{(\theta, \beta) \in N} \|\varepsilon_t(\theta, \beta)\|_r, \sup_t \sup_{(\theta, \beta) \in N} \left\| \frac{\partial \varepsilon_t(\theta, \beta)}{\partial(\theta, \beta)} \right\|_r \leq K$ for some finite constant K .

Denote P^* to be probability distribution of the generated samples $y_{s+\tau}^*$ conditioning on $\{y_{s+\tau}, x_s\}_{s=1, \dots, T+P-\tau}$. Denote \Rightarrow^* to be "weak convergence" in P^* to distinguish weak convergence in the original probability measure (\Rightarrow). The following theorem validates the consistency of the bootstrap approximation of the distribution of the test statistic \mathcal{T}_{LRT} .

Theorem 2 Assume Assumptions 1 – 5. Then,

$$\mathcal{T}_{LRT}^* \Rightarrow^* \max \{ \mu_1(\mathcal{F}, \mathcal{V})' \mathcal{V}^{-1} \mu_1(\mathcal{F}, \mathcal{V}), \dots, \mu_K(\mathcal{F}, \mathcal{V})' \mathcal{V}^{-1} \mu_K(\mathcal{F}, \mathcal{V}) \}.$$

3.3 Use of Normal Approximation

Notice that the limit distribution in Theorem 1 is nonstandard and a complicated functional of Brownian motion. This is mainly because when the forecasting models are nested and $\lim_{P, T \rightarrow \infty} P/T = \lambda$, $\lambda > 0$, the limiting distribution of $\bar{f}_m^{adj} / \sqrt{\hat{v}_m}$ becomes a functional of Brownian motion, instead of a standard normal distribution. (e.g., Clark and McCracken (2001 and 2005), McCracken (2007)).

However, for pairwise model comparison and under a different set of assumptions asymptotic validity of the normal distribution has been shown in Giacomini and White (2006)¹⁴ and approximate asymptotic validity of the normal distribution follows from Clark and McCracken (2001, 2005) (see HW for a more detailed discussion). The asymptotics and arguments for the validity of the normal approximation are taken as justification for the assumption of normality of $\bar{f}_m^{adj} / \sqrt{\hat{v}_m}$ in HW where inference is based on the maximum of correlated normals, building on results from the literature of order statistics. The simulation

¹⁴Note that we reference the technical conditions and not the procedures of Giacomini and White (2006).

experiments in CW and HW provide evidence for one-step ahead forecasts and homoskedastic prediction errors that the size properties of their test statistics with the standard normal approximation of $\bar{f}_m^{adj}/\sqrt{\hat{v}_m}$ are reasonable. They also demonstrate that the standard normal approximation performs reasonably well in heteroskedastic environments when the number of additional regressors, k_m is equal to one.¹⁵ Moreover, they do not find substantial size or power improvements when using simulated or bootstrapped critical values rather than asymptotic normal critical values.

Based on the results in CW and HW, one may conjecture that treating \bar{f}^{adj} as normal might deliver a reasonable approximation despite the limiting distribution of the Likelihood Ratio test being a functional of Brownian motion. This leads us to use the critical values computed under the normality assumption for the MSPE-adjusted to evaluate the tests. In order to compute the critical values we proceed as follows.

Treat $\bar{f}^{adj} \sim N(0, V)$. Define $Z = (Z_1, \dots, Z_M)' \sim N\left(0, \hat{R}^{1/2}\hat{V}\hat{R}^{1/2}\right)$, where $\hat{R} = \text{diag}(\hat{V})$. Then, we approximate the distributions of the tests as follows: $\chi^2(M)$ for the limit of \mathcal{T}_{chi} , $\max_{1 \leq m \leq M} \{Z_1, \dots, Z_M\}$ for the limit of $\mathcal{T}_{max t}$, while for the limit of \mathcal{T}_{LRT} $\max_{1 \leq k \leq K} \left\{ \mu_1 \left(Z, \hat{V} \right)' \hat{V}^{-1} \mu_1 \left(Z, \hat{V} \right), \dots, \mu_K \left(Z, \hat{V} \right)' \hat{V}^{-1} \mu_K \left(Z, \hat{V} \right) \right\}$.

4 Monte Carlo Simulation

We now outline in detail the two experimental designs for the Monte Carlo simulation: one motivated by empirical studies on the predictive content of the yield curve for gdp growth and one suited to the comparison of forecast models for inflation. The evaluation of the tests is implemented with asymptotic critical values derived through simulations for two settings: first based on the assumption of normality of the adjusted MSPE and second through bootstrap. Because the tests, as we will discuss shortly, are subject to size distortions we also report results for the size-adjusted power of the tests.

4.1 Experimental Design

The implementation of the simulation exercise requires the design of the DGP process for the size and the power experiment and the selection of the forecasting models.

¹⁵The simulation results in CW for pairwise model comparison show an empirical size between 5% and 10% for a 10% nominal size for both heteroskedastic and homoskedastic forecast errors, for both the expanding window and rolling estimation scheme and for values of λ ranging between one third and six.

The design for DGP1 takes the form:

$$y_{t+\tau} = c + \rho y_t + \gamma' x_{t+1} + u_{t+\tau} \quad (6)$$

with $c = 1$, $\rho = 0.25$ and $u_{t+\tau}$ i.i.d $N(0, 1)$ when $\tau = 1$ and $u_{t+\tau}$ a $MA(\tau - 1)$ process of the form:¹⁶ $u_{t+\tau} = \varepsilon_{t+\tau} + 0.95\varepsilon_{t+\tau-1} + 0.9\varepsilon_{t+\tau-2} + 0.8\varepsilon_{t+\tau-3}$ when $\tau = 4$. The DGP for the size exercise is an autoregressive process obtained by setting $\gamma = \mathbf{0}$. In the power experiment three exogenous variables are added to the autoregressive term and $\gamma = [0.05 \ 0.05 \ 0.25]'$ when $\tau = 1$ and $\gamma = [0.15 \ 0.15 \ 0.75]'$ when $\tau = 4$. The exogenous variables, collected in the vector $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$ are determined independently by:

$$x_{i,t} = a_i + \delta_i x_{i,t-1} + \nu_{it} \quad (7)$$

with $a_i = 1$, $\delta_i = 0.8$ for $i = 1, 2, 3$, $\nu_{it} \sim N(0, 1)$, ν_{it} is independent of u_t and of ν_{jt} for all $i \neq j$ and t . DGP1 is loosely related to the empirical analysis conducted in Ang et al. (2006). The variable of interest for forecasting is gdp growth which exhibits little persistence. The exogenous variables are quite persistent, as it is the case for short term bonds yield, bonds spread and inflation, the candidate variables for predicting gdp growth. The vector γ is chosen such that one of the variables matters much more than the others in determining the evolution of y_t . This is consistent with the findings in Ang et al. (2006) that short rate and lagged gdp growth account five times less than the term spread in an estimated univariate and unconstrained linear regression model for gdp growth.

Similarly to DGP1 the target series for DGP2 $y_{t+\tau}$ is generated by the process:

$$y_{t+\tau} = c + \rho y_t + \gamma' x_t + u_{t+\tau}.$$

The vector x_t follows a VAR(1) process:

$$\mathbf{x}_t = \mathbf{a} + \Phi \mathbf{x}_{t-1} + \mathbf{v}_t$$

with $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$, \mathbf{a} a 3×1 vector of ones, $u_{t+\tau}$ follows a $MA(\tau - 1)$ process of the same form as for DGP1, $\mathbf{v}_t \sim N(0, I)$ and $u_{t+\tau}$ is independent of \mathbf{v}_t . The VAR(1) regression

¹⁶The MA process for the errors is taken from Clark and McCracken (2011).

coefficient Φ allows for interdependence between the predictors:

$$\Phi = \begin{bmatrix} 0.6 & 0.1 & 0 \\ 0.6 & 0.25 & 0 \\ 0 & 0 & 0.9 \end{bmatrix}. \quad (8)$$

In the size exercise the vector γ is set to zeros, while in the power exercise $\gamma = [0.05 \ 0.05 \ -0.05]'$ for one step ahead forecasts and $\gamma = [0.25 \ 0.25 \ -0.25]'$ for four step ahead forecasts. The second design is based on the empirical application presented in this paper, where the US CPI core inflation is forecasted with a backward looking Phillips Curve type models using a recessionary gap variable as in Stock and Watson (2010) and inflation components as in Hubrich (2005) and Hendry and Hubrich (2011). The coefficient matrix Φ is obtained by estimating a VAR(1) for cpi food inflation, cpi energy inflation, and a recessionary gap defined in Section 5 over the sample 1959:Q1-2010Q2. The gap evolves independently of the two components and in the power exercise it is negatively correlated with core inflation.

Next the forecasting models are selected. The benchmark model is the true model in the size experiment:

$$M_0 : y_{t+\tau} = c_0 + \beta_0 y_t + u_{0,t+\tau}. \quad (9)$$

while the alternative models take the form:

$$M_m : y_{t+\tau} = \beta_m' x_{m,t+1} + u_{m,t+\tau}, \quad m = 1, \dots, M, \quad (10)$$

where for DGP1 $x_{m,t+1} = (1, y_t, x_{1,t+1})'$ for $m = 1$, $x_{m,t+1} = (1, y_t, x_{1,t+1}, x_{2,t+1})'$ for $m = 2$, $x_{m,t+1} = (1, y_t, x_{1,t+1}, x_{2,t+1}, x_{3,t+1})'$ for $m = 3$, while for DGP2 $x_{m,t+1} = (1, y_t, x_{1,t})'$ for $m = 1$, $x_{m,t+1} = (1, y_t, x_{1,t}, x_{2,t})'$ for $m = 2$, $x_{m,t+1} = (1, y_t, x_{1,t}, x_{2,t}, x_{3,t})'$ for $m = 3$. The dimension of β_m is $m + 2$. Model M3 then nests not only the benchmark but also models 1 and 2; this is equivalent to the scenario analyzed in the first case presented in Section 2.1, where the alternative models are progressively nested within each other.

For both DGPs the estimates are carried out through OLS, with rolling and expanding window scheme, for 5 and 10 percent significance level. The results under the two schemes and for the two significance levels are qualitatively similar so for brevity we report only the ones for the expanding window scheme and for 10% significance level.¹⁷ We focus on one and four step ahead forecasts and we consider different lengths of the in-sample, $T =$

¹⁷Results for the rolling scheme and for 5% significance level are available from the authors upon request.

$\{40, 80, 100, 200\}$, and out-of-sample $P = \{40, 100\}$. The sample sizes selected are consistent with the current length of time series available at the quarterly frequency.

4.2 Simulation Results

First we report the empirical size of the tests for the chi-squared, the max-t stat and the LRT tests described in Section 2.1 and 2.2 for data generated from both DGPs. Table 1 provides empirical size results for critical values derived through bootstrapping while Table 2 shows the empirical size for critical values obtained by relying on the normal approximation.

Table 1. Empirical Size Bootstrapped Critical Values DGP1 and DGP2 (Nominal Size =10%)

		DGP1								DGP2							
		$\tau = 1$															
test	T	P=40				P=100				P=40				P=100			
		40	80	100	200	40	80	100	200	40	80	100	200	40	80	100	200
chi ²		0.088	0.112	0.094	0.103	0.091	0.089	0.099	0.126	0.107	0.103	0.105	0.104	0.100	0.122	0.120	0.110
max-t		0.085	0.079	0.087	0.085	0.076	0.087	0.069	0.092	0.124	0.117	0.131	0.118	0.126	0.117	0.106	0.110
LRT _I		0.092	0.102	0.092	0.102	0.077	0.086	0.089	0.104	0.095	0.117	0.109	0.099	0.099	0.121	0.090	0.104
LRT _D		0.092	0.102	0.092	0.102	0.077	0.086	0.089	0.104	0.105	0.110	0.126	0.105	0.125	0.097	0.107	0.094
chi ² -unadj		0.079	0.098	0.089	0.105	0.083	0.081	0.094	0.114	0.105	0.100	0.097	0.110	0.107	0.117	0.130	0.114
max-t-unadj		0.092	0.089	0.108	0.089	0.102	0.090	0.078	0.091	0.127	0.121	0.139	0.132	0.130	0.137	0.112	0.118
LRT _I -unadj		0.089	0.101	0.098	0.102	0.084	0.087	0.088	0.103	0.088	0.116	0.105	0.114	0.083	0.108	0.098	0.100
LRT _D -unadj		0.089	0.101	0.098	0.102	0.084	0.087	0.088	0.103	0.103	0.120	0.134	0.102	0.119	0.095	0.088	0.099
		$\tau = 4$															
chi ²		0.183	0.193	0.215	0.235	0.188	0.207	0.193	0.210	0.110	0.131	0.115	0.126	0.108	0.112	0.104	0.124
max-t		0.131	0.157	0.132	0.158	0.135	0.140	0.120	0.138	0.151	0.147	0.135	0.116	0.117	0.132	0.102	0.148
LRT _I		0.152	0.186	0.176	0.208	0.151	0.169	0.152	0.169	0.116	0.126	0.121	0.129	0.127	0.118	0.106	0.138
LRT _D		0.152	0.186	0.176	0.208	0.151	0.169	0.152	0.169	0.114	0.118	0.134	0.128	0.121	0.130	0.098	0.125
chi ² -unadj		0.159	0.186	0.196	0.218	0.173	0.211	0.205	0.199	0.101	0.117	0.107	0.117	0.105	0.108	0.116	0.121
max-t-unadj		0.133	0.154	0.120	0.155	0.125	0.126	0.105	0.122	0.148	0.133	0.131	0.120	0.116	0.130	0.102	0.151
LRT _I -unadj		0.155	0.177	0.153	0.199	0.138	0.166	0.152	0.165	0.119	0.125	0.123	0.114	0.115	0.113	0.103	0.133
LRT _D -unadj		0.155	0.177	0.153	0.199	0.138	0.166	0.152	0.165	0.127	0.128	0.131	0.119	0.120	0.119	0.098	0.129

NOTE: The DGPs are described in Section 4.1. T and P refers to the size of the in-sample and out-of-sample respectively. The forecast horizon is denoted by τ . The suffix '-unadj' refers to test statistics constructed based on the differences in MSPE without using the CW adjustment. The reported results are based on 10000 Monte Carlo draws when the statistics are evaluated against the empirical critical values or when critical values are obtained under the assumption of normality. For critical values generated through the bootstrap the number of draws is 1000 and for each draw we generate 500 bootstrap samples. For every draw, the initial 100 observations generated are discarded.

For one step ahead the tests based on the fixed regressor bootstrap show reasonable finite sample size properties with empirical size ranging from 6.9 to 12.6 for DGP1 and between 8.3 and 13.9 for DGP2 and median of 9.1 and 10.9, respectively. For four step ahead forecasts the

size deteriorates with respect to one-step ahead forecasts in the case of DGP1; in particular all tests tend to over-reject and the size distortions are larger for the chi-squared test. For DGP2 instead the good size properties are preserved for the longer forecast horizon. From Table 1 it also emerges that size does not vary systematically with either T or P or with the CW adjustment.¹⁸

For critical values derived by treating the MSPE differential as normally distributed the rejection rates depend on the incorporation of the CW adjustment: for one step ahead forecasts the tests based on MSPE-adjusted are generally well-sized confirming the results obtained by HW and Clark and McCracken (2011). Deviations from nominal size are broadly similar to the ones obtained with the fixed regressors bootstrap, the empirical size in the upper panel of Table 2 for the max-t and the LRT_D tests ranging from 6.5 and 12.0 for DGP1 and from 6.5 and 12.8 for DGP2 for a 10% nominal size.

Table 2. Empirical Size under Normality DGP1 and DGP2 (Nominal Size =10%)

		DGP1								DGP2							
		$\tau = 1$															
test	T	P=40				P=100				P=40				P=100			
		40	80	100	200	40	80	100	200	40	80	100	200	40	80	100	200
chi ²		0.145	0.150	0.144	0.156	0.114	0.121	0.128	0.135	0.147	0.146	0.150	0.154	0.111	0.123	0.121	0.136
max-t		0.115	0.120	0.118	0.125	0.095	0.100	0.099	0.101	0.128	0.127	0.122	0.128	0.101	0.098	0.106	0.110
LRT_I		0.127	0.136	0.131	0.135	0.092	0.101	0.101	0.110	0.130	0.126	0.129	0.133	0.097	0.097	0.098	0.112
LRT_D		0.093	0.101	0.097	0.107	0.065	0.071	0.066	0.069	0.099	0.095	0.101	0.104	0.066	0.065	0.066	0.074
chi ² -unadj		0.189	0.167	0.161	0.167	0.200	0.174	0.176	0.163	0.179	0.168	0.163	0.165	0.182	0.175	0.166	0.159
max-t-unadj		0.026	0.046	0.049	0.070	0.008	0.019	0.023	0.037	0.032	0.047	0.053	0.079	0.007	0.017	0.023	0.040
LRT_I -unadj		0.071	0.090	0.092	0.105	0.044	0.055	0.060	0.072	0.070	0.083	0.089	0.103	0.039	0.053	0.055	0.072
LRT_D -unadj		0.026	0.043	0.045	0.063	0.006	0.013	0.017	0.026	0.026	0.037	0.046	0.063	0.005	0.011	0.014	0.026
		$\tau = 4$															
chi ²		0.393	0.398	0.394	0.400	0.283	0.271	0.269	0.270	0.366	0.362	0.353	0.349	0.230	0.237	0.235	0.252
max-t		0.236	0.238	0.234	0.238	0.175	0.166	0.169	0.179	0.232	0.226	0.216	0.217	0.168	0.157	0.167	0.167
LRT_I		0.335	0.340	0.338	0.339	0.231	0.216	0.214	0.224	0.317	0.304	0.292	0.284	0.197	0.185	0.192	0.199
LRT_D		0.279	0.276	0.277	0.278	0.172	0.161	0.159	0.173	0.269	0.258	0.246	0.242	0.156	0.142	0.154	0.159
chi ² -unadj		0.430	0.410	0.410	0.402	0.337	0.330	0.324	0.295	0.374	0.369	0.363	0.353	0.283	0.274	0.278	0.269
max-t-unadj		0.080	0.120	0.124	0.154	0.021	0.040	0.045	0.082	0.087	0.113	0.119	0.145	0.025	0.041	0.047	0.080
LRT_I -unadj		0.217	0.252	0.260	0.284	0.100	0.125	0.130	0.158	0.214	0.225	0.227	0.239	0.100	0.112	0.113	0.138
LRT_D -unadj		0.131	0.167	0.177	0.205	0.027	0.048	0.052	0.082	0.132	0.155	0.156	0.177	0.029	0.044	0.046	0.076

NOTE: Refer to the note on Table 1.

¹⁸This last finding was expected given that the bootstrap should take care of the recentering of the statistics

The chi-squared test is always oversized, as found in HW, while LRT_I exhibits slightly more size distortions than max-t and LRT_D . Also from Table 2 the behaviour of max-t and LRT_D changes with P and T and in general for a fixed T size decreases with P. In the case of four step ahead forecasts the tests are grossly oversized for small P, although the size distortions decrease considerably as the out-of-sample size increases. For both DGPs the max-t stat and the LRT_D have empirical size closer to the nominal size. For example, for T=100 the empirical size ranges from 14 to 18%.¹⁹ For tests constructed using the MSPE-unadjusted the discrepancies between nominal and empirical size are quite large for both DGPs and both forecast horizons: the chi-squared-unadjusted over-rejects, while the other tests generally under-reject with the LRT_D -unadjusted suffering the greatest distortions, the empirical size being even 20 times smaller than the nominal size.

Table 3. Power, Bootstrapped Critical Values DGP1 and DGP2

		DGP1								DGP2							
		$\tau = 1$															
test	T	P=40				P=100				P=40				P=100			
		40	80	100	200	40	80	100	200	40	80	100	200	40	80	100	200
chi ²		0.356	0.388	0.405	0.410	0.811	0.858	0.835	0.860	0.350	0.335	0.359	0.370	0.690	0.659	0.675	0.610
max-t		0.416	0.474	0.481	0.498	0.847	0.894	0.884	0.903	0.494	0.456	0.451	0.445	0.680	0.645	0.641	0.583
LRT_I		0.385	0.439	0.453	0.461	0.847	0.881	0.856	0.870	0.381	0.360	0.375	0.375	0.702	0.667	0.680	0.610
LRT_D		0.385	0.439	0.453	0.461	0.847	0.881	0.856	0.870	0.474	0.433	0.434	0.417	0.749	0.715	0.721	0.639
chi ² -unadj		0.119	0.143	0.140	0.140	0.235	0.330	0.273	0.287	0.160	0.208	0.201	0.229	0.279	0.365	0.374	0.371
max-t-unadj		0.336	0.321	0.321	0.299	0.700	0.688	0.633	0.572	0.404	0.342	0.377	0.357	0.558	0.534	0.484	0.435
LRT_I -unadj		0.198	0.194	0.200	0.200	0.431	0.454	0.432	0.414	0.219	0.242	0.219	0.243	0.337	0.393	0.400	0.380
LRT_D -unadj		0.198	0.194	0.200	0.200	0.431	0.454	0.432	0.414	0.350	0.341	0.312	0.311	0.560	0.536	0.517	0.462
		$\tau = 4$															
chi ²		0.441	0.494	0.483	0.489	0.872	0.910	0.916	0.908	0.241	0.238	0.242	0.263	0.464	0.488	0.460	0.524
max-t		0.493	0.562	0.556	0.559	0.882	0.931	0.924	0.940	0.425	0.455	0.431	0.486	0.658	0.670	0.676	0.732
LRT_I		0.483	0.525	0.539	0.538	0.887	0.917	0.918	0.938	0.284	0.293	0.315	0.314	0.560	0.560	0.557	0.607
LRT_D		0.483	0.525	0.539	0.538	0.887	0.917	0.918	0.938	0.360	0.393	0.410	0.404	0.642	0.651	0.667	0.738
chi ² -unadj		0.148	0.193	0.191	0.205	0.119	0.162	0.170	0.239	0.133	0.144	0.146	0.176	0.108	0.171	0.141	0.213
max-t-unadj		0.214	0.238	0.244	0.248	0.299	0.359	0.357	0.409	0.368	0.392	0.365	0.377	0.567	0.556	0.529	0.563
LRT_I -unadj		0.175	0.214	0.209	0.219	0.184	0.245	0.229	0.306	0.225	0.234	0.223	0.231	0.261	0.293	0.265	0.330
LRT_D -unadj		0.175	0.214	0.209	0.219	0.184	0.245	0.229	0.306	0.319	0.319	0.329	0.324	0.527	0.555	0.507	0.560

NOTE: Refer to the note on Table 1.

¹⁹Further research might also explore possibilities of improving the size of the tests for multiple model comparisons based on the normal approximation using different HAC estimators for the variance of \hat{f}^{adj} and \hat{f}^{unadj} than the one proposed by Newey and West (1987). This might be a promising research avenue given that Clark and McCracken (2011b) for pairwise model comparison find considerable size improvements for certain other HAC estimators.

We now comment on the power of the tests given the results in Table 3 through Table 5. In the case of a multivariate one-sided alternative hypothesis there is no uniformly more powerful test, so the ranking between tests should vary across different simulation designs. However we expect the chi-squared test to have the lowest power as it disregards the one sided nature of the alternative. Also, given the structure of the alternative models the LRT_D test should outperform the LRT_I test as the latter does not account for the particular ordering of the MSPE differentials. These conjectures are confirmed by our simulations.

Table 4. Power under Normality DGP1 and DGP2

		DGP1								DGP2							
		$\tau = 1$															
test	T	P=40				P=100				P=40				P=100			
		40	80	100	200	40	80	100	200	40	80	100	200	40	80	100	200
chi ²		0.535	0.578	0.583	0.593	0.898	0.918	0.927	0.941	0.567	0.589	0.619	0.620	0.910	0.937	0.948	0.947
max-t		0.604	0.666	0.691	0.704	0.927	0.956	0.960	0.977	0.800	0.839	0.840	0.873	0.979	0.988	0.994	0.992
LRT_I		0.571	0.613	0.620	0.629	0.917	0.936	0.942	0.958	0.665	0.701	0.703	0.733	0.946	0.965	0.972	0.975
LRT_D		0.614	0.660	0.674	0.684	0.938	0.954	0.955	0.970	0.713	0.758	0.762	0.797	0.963	0.979	0.985	0.984
chi ² -unadj		0.202	0.233	0.233	0.245	0.335	0.382	0.399	0.418	0.188	0.237	0.246	0.288	0.253	0.359	0.381	0.455
max-t-unadj		0.143	0.239	0.261	0.326	0.265	0.407	0.450	0.534	0.337	0.447	0.469	0.535	0.543	0.672	0.695	0.759
LRT_I -unadj		0.153	0.209	0.221	0.253	0.283	0.362	0.393	0.442	0.249	0.327	0.339	0.391	0.360	0.484	0.502	0.586
LRT_D -unadj		0.164	0.229	0.247	0.283	0.320	0.409	0.440	0.498	0.278	0.375	0.394	0.455	0.428	0.573	0.593	0.673
		$\tau = 4$															
chi ²		0.723	0.753	0.767	0.767	0.924	0.947	0.951	0.961	0.521	0.531	0.528	0.537	0.614	0.633	0.664	0.677
max-t		0.702	0.738	0.772	0.781	0.933	0.959	0.966	0.980	0.558	0.587	0.614	0.629	0.722	0.763	0.792	0.808
LRT_I		0.717	0.749	0.766	0.769	0.928	0.948	0.954	0.966	0.552	0.571	0.567	0.586	0.666	0.689	0.720	0.739
LRT_D		0.746	0.774	0.789	0.796	0.940	0.960	0.963	0.973	0.566	0.588	0.603	0.620	0.701	0.736	0.767	0.788
chi ² -unadj		0.468	0.480	0.479	0.486	0.504	0.549	0.557	0.577	0.375	0.391	0.386	0.405	0.240	0.287	0.299	0.356
max-t-unadj		0.280	0.383	0.402	0.477	0.364	0.493	0.539	0.626	0.262	0.327	0.359	0.424	0.263	0.372	0.393	0.494
LRT_I -unadj		0.349	0.406	0.415	0.446	0.403	0.482	0.503	0.557	0.331	0.373	0.386	0.427	0.248	0.318	0.344	0.412
LRT_D -unadj		0.349	0.413	0.427	0.465	0.426	0.505	0.533	0.595	0.313	0.370	0.395	0.456	0.257	0.352	0.379	0.472

NOTE: Refer to the note on Table 1.

Results in Table 3 and Table 4 show that for all tests and forecast horizons the power increases with the size of the out-of-sample for a given in-sample size. Improvements are also generally obtained when the in-sample grows for fixed out of sample but the power gain is more substantial for an increase in P than an equal increase in T. Power is larger when the critical values are obtained under the assumption of normality than for bootstrapped critical values.²⁰ For tests based on the MSPE-adjusted the performance of the chi-squared test is

²⁰Results from Clark and McCracken (2011) show opposite behavior of the power properties.

overall disappointing: as expected it always ranks last for both DGPs and for both horizons. The LRT_I ranks third and the max-t and LRT_D have comparable performance but their ranking changes with P and T. The same ranking applies to the tests based on the MSPE unadjusted with a few exceptions in the case of critical values derived under normality but this might be related to the size distortions documented in Table 2. In general the tests based on MSPE-adjusted have higher power than the tests based on MSPE.

Table 5. Size-Adjusted Power DGP1 and DGP2

		DGP1								DGP2							
		$\tau = 1$															
test	T	P=40				P=100				P=40				P=100			
		40	80	100	200	40	80	100	200	40	80	100	200	40	80	100	200
chi ²		0.447	0.492	0.493	0.511	0.880	0.906	0.908	0.923	0.492	0.493	0.523	0.532	0.904	0.921	0.933	0.935
max-t		0.569	0.630	0.640	0.652	0.931	0.956	0.961	0.976	0.747	0.798	0.814	0.841	0.979	0.988	0.994	0.991
LRT_I		0.521	0.556	0.554	0.570	0.923	0.935	0.941	0.954	0.602	0.630	0.674	0.675	0.947	0.963	0.973	0.975
LRT_D		0.631	0.667	0.672	0.691	0.959	0.969	0.973	0.981	0.718	0.759	0.793	0.797	0.976	0.986	0.991	0.991
chi ² -unadj		0.113	0.149	0.157	0.158	0.206	0.265	0.281	0.320	0.114	0.160	0.171	0.199	0.169	0.255	0.282	0.345
max-t-unadj		0.418	0.440	0.443	0.421	0.823	0.800	0.812	0.784	0.596	0.614	0.619	0.599	0.928	0.928	0.914	0.894
LRT_I -unadj		0.198	0.223	0.239	0.244	0.444	0.482	0.500	0.506	0.311	0.365	0.366	0.385	0.537	0.622	0.636	0.658
LRT_D -unadj		0.393	0.401	0.396	0.380	0.822	0.773	0.782	0.753	0.551	0.591	0.568	0.575	0.901	0.909	0.909	0.880
		$\tau = 4$															
chi ²		0.316	0.345	0.345	0.365	0.657	0.792	0.817	0.847	0.200	0.192	0.202	0.215	0.428	0.452	0.478	0.486
max-t		0.381	0.418	0.429	0.444	0.681	0.853	0.876	0.923	0.332	0.361	0.389	0.408	0.600	0.669	0.707	0.718
LRT_I		0.360	0.388	0.392	0.418	0.704	0.832	0.859	0.888	0.238	0.243	0.253	0.284	0.505	0.548	0.581	0.606
LRT_D		0.436	0.469	0.482	0.500	0.775	0.888	0.907	0.933	0.290	0.302	0.322	0.361	0.607	0.652	0.684	0.716
chi ² -unadj		0.125	0.144	0.142	0.154	0.229	0.269	0.282	0.315	0.120	0.128	0.135	0.151	0.100	0.125	0.140	0.179
max-t-unadj		0.343	0.330	0.335	0.332	0.721	0.738	0.734	0.690	0.290	0.298	0.321	0.340	0.542	0.567	0.563	0.555
LRT_I -unadj		0.184	0.191	0.197	0.200	0.403	0.425	0.449	0.457	0.182	0.190	0.208	0.225	0.249	0.298	0.319	0.343
LRT_D -unadj		0.290	0.287	0.278	0.285	0.678	0.662	0.663	0.639	0.257	0.276	0.287	0.308	0.497	0.528	0.550	0.535

NOTE: Refer to the note on Table 1.

However, the analysis of the results for the size properties of the tests highlighted some size distortions depending on the horizon, the statistic and the critical values used. For this reason we provide with size-adjusted power results in Table 5. In this case the ranking of the tests using MSPE-adjusted varies only with the DGP: LRT_D ranks first for DGP1 while max-t ranks first for DGP2, LRT_I ranks always third and the chi-squared always ranks last. The tests based on MSPE-unadjusted are again clearly outperformed in terms of power by their respective counterparts based on MSPE-adjusted.

Summing up, for the case of bootstrapped critical values all tests are generally well behaved in terms of size regardless of the forecast horizon considered and regardless of the recentering of the sample MSPE. Relying on the normal approximation to obtain critical values yields actual sizes close to the nominal sizes for one step ahead forecasts and for tests based on MSPE-adjusted.²¹ For the power exercise recentering the MSPE as suggested by CW and HW increases power; assuming normality of the MSPE differentials yields higher power than bootstrapping. The chi-squared test performs poorly, followed by the LRT_I ; LRT_D and max-t rank first where the ranking between LRT_D and max-t depends on the simulation setting.

5 Forecasting US Inflation

In this section we apply our test to the evaluation of equal predictive ability for forecasting the US CPI core yearly inflation rate. Inflation exhibits very different characteristics over the last 50 years: in the beginning of the sample it is very high and volatile while from the mid-80s it is more stable and has a lower mean. This led us to split the data into two samples, as the different behavior is possibly due to parameters instability²²not handled by our framework: the first includes the observations 1959:Q1-1971:Q4, the second spans from 1984:Q1 through 1997:Q4. The remaining years (1972-1983 for the first sample, and 1998-2010 for the second sample) are used for forecast evaluation. For the same value of the in-sample ($T=51$ or 56) and similar values of the out-of-sample ($P=48$ or $P=50$) the simulations results summarized in Section 4 suggest that the tests are well behaved in terms of size and power. The models we consider are an AR(1) benchmark with constant and three alternatives obtained by progressively expanding the set of predictors: a lagged real activity gap measure in model M1, the lagged inflation rate for cpi food in model M2 and lagged inflation rate for cpi energy in model M3. The real activity gap we consider is the recessionary gap defined by Stock and Watson (2010) as the difference between the current unemployment rate and the minimum unemployment rate over the current and previous 11 quarters. Stock and Watson (2010) show evidence of a linear relationship between PCE inflation and the recessionary gap, a finding that is relevant for our backward Phillips-curve type of analysis for core CPI inflation. Food and energy inflation are the two most volatile

²¹This result was emphasized for the pairwise model comparison of nested models by Clark and West (2006, 2007) and for multimodel comparison in HW.

²²See for example Boivin and Giannoni (2002), Cogley, T. and T.J. Sargent (2001).

components of CPI all items inflation and are excluded from the computation of CPI core inflation. We ask whether those two components have any additional predictive ability over a Phillips Curve model for CPI core inflation.²³ Consistently with the simulation settings, all the alternative models nest the benchmark and the alternative models are nested within each other. The estimation technique adopted is recursive OLS applied to the annualized quarterly inflation rate and the forecast horizons of interest are one and four.

Table 6 collects test results for the chi-squared, the max t-statistic and the two likelihood ratio type tests for both one step ahead and four step ahead forecasts.²⁴

Table 6. Test of Equal Forecast Accuracy for US Inflation

	<i>One – Step Ahead</i>						<i>Four – Step Ahead</i>					
	1st sample			2nd sample			1st sample			2nd sample		
	test stat	p-values		test stat	p-values		test stat	p-values		test stat	p-values	
		B	N		B	N		B	N		B	N
chi ²	4.301	0.631	0.231	2.486	0.636	0.478	4.348	0.226	0.250	2.029	0.672	0.566
max-t	1.883	0.659	0.084	1.484	0.482	0.181	2.922	0.032	0.026	1.687	0.148	0.158
LRT _I	4.300	0.544	0.121	2.487	0.591	0.340	4.349	0.110	0.146	2.020	0.402	0.439
LRT _D	4.301	0.537	0.078	0.712	0.633	0.538	4.301	0.098	0.123	0.143	0.640	0.600

NOTE: The variable to be forecasted is the annualized quarter to quarter inflation rate for PCE core. The models we consider are an AR(1) benchmark with constant (M0) and three alternatives obtained by progressively expanding the set of predictors: a lagged real activity gap measure in model M1, the lagged inflation rate for cpi food in model M2 and lagged inflation rate for cpi energy in model M3. The estimation samples are 1959:Q1-1971:Q4 and 1984:Q1 through 1997:Q4. The remaining years (1972-1983 for the first sample, and 1998-2010 for the second sample) are used for forecast evaluation.

For each sample the table reports the test statistics and the p-values obtained under the assumption of Normality (N) of the MSPE and through bootstrapping (B). For the second sample there is clear evidence from all tests that equal predictability at both forecast horizons cannot be rejected. For the first sample instead the results are mixed: at one step ahead when the normal approximation is used both max-t stat and LRT_D reject the null at the 10% significance level, while the recessionary gap and/or the food and/or energy components do not have significant predictive content for core inflation when critical values

²³Hubrich (2005) and Hendry and Hubrich (2011) discuss the merit of including components in the forecasting model for the aggregate; the latter authors particularly suggest to include components in the forecasting model for the aggregate.

²⁴Given the bad size and power performance of the unadjusted statistics we only report results for the tests based on the CW adjustment.

are bootstrapped.²⁵ This finding is consistent with the higher power of the max-t test and LRT_D under the normal approximation for a 1 step-ahead horizon documented in the upper panel of Table 4. For four step-ahead forecasts there is strong evidence against the null for the max-t stat and the LRT-type tests for bootstrapped critical values, while for critical values under normality only the max-t rejects. This is consistent with our simulation findings of higher power of those tests based on bootstrapped critical values for four step ahead forecasts.²⁶ We take our simulation results as evidence in favour of applying the max-t test and LRT_D under the normal approximation for one step ahead forecasts, and the max-t test and LRT_D with bootstrapped critical values for a four step ahead forecast horizon. In light of these considerations we conclude that we reject equal forecast accuracy for the 1st sample on a 10% nominal significance level.

6 Conclusions

This paper introduces a likelihood ratio type predictability test for the comparison of a small number of models nesting a parsimonious benchmark model. In formulating the alternative hypothesis and the test statistics we distinguish among three cases according to the structure of the alternative models. We show that the limiting distribution of the test statistic is non-standard and it depends on the characteristics of the predictors. Then we prove the validity of the bootstrap procedure developed in Clark and McCracken (2011) for our proposed test.

The finite sample size and power properties of the test are evaluated via Monte Carlo simulations either by treating the statistics as normally distributed or by bootstrapping. These investigations indicate that the normal approximation of the vector of MSPE-adjusted yields approximately correctly sized tests for one step ahead forecasts but for longer horizons and for small out-of-sample size it provides grossly oversized tests, even though size improves with increasing out-of-sample size. Further research might explore possibilities of improving the size of the tests based on the normal approximation using different HAC estimators. Instead the bootstrapped critical values deliver tests that have empirical size close to nominal size also for longer forecast horizons. Relying on the bootstrap rather than on the assumption of normality to compute the critical values can negatively affect the power of the tests. Also,

²⁵The discrepancies between the p-values based on the normal and bootstrap critical values seem to be due to small sample issues as explored in additional simulation experiments.

²⁶Concerns may arise on the stability of the parameters during the last recession, so we repeat the analysis for the second sample disregarding the observations past 2007Q2. For this shorter sample for both forecast horizons all tests fail to reject the null regardless of the methods under which the critical values are obtained.

we find that the CW adjustment improves power. We compare our test with two existing tests (chi-squared and max-t statistic tests as in HW) and we find that while the chi-square test performs poorly in terms of power the LR-type test and the max t-statistic have better, comparable power properties but the ranking depends on parameterization of the Monte Carlo experiment.

Last, in the empirical analysis we find that the recessionary gap and the food and energy components do not have predictive content for core inflation during the Great Moderation period while the tests provide mixed evidence in the earlier sample. Therefore, conclusions on the predictive ability of a Phillips type curve for US core inflation depend not only on the sample, but also on the test and on the method with which the critical values are obtained. However, the size and power performance of the tests outlined in the simulation results can provide guidance on which test and critical values are more reliable in this environment.

References

- [1] Andersson, M., D'Agostino, A., de Bondt, G. J. and R. Moreno (2011), 'The Predictive Content of Sectoral Stock Prices: A US-Euro Area Comparison', ECB wp. 1343.
- [2] Ang, A., Piazzesi, M., and M. Wei (2006), 'What does the Yield Curve tell us about GDP growth?', *Journal of Econometrics*, vol.131, 359-403.
- [3] Ashley, R., C.W.J. Granger and R.Schmalense (1980), 'Advertising and Aggregate Consumption: an Analysis of Causality', *Econometrica*, vol.48 n.5.
- [4] Boivin, J. and M. Giannoni (2002), 'Assessing Changes in the Monetary Transmission Mechanism: A VAR Approach', *Federal Reserve Bank of New York Monetary Policy Review*, May, 97-111.
- [5] Chao, J. C., Corradi V. and Norman R. Swanson (2001), 'An Out-of-Sample Test for Granger Causality', *Macroeconomic Dynamics* vol.5, 598-620.
- [6] Clark, T.E. and M. W. McCracken (2000), 'Not-for-Publication Appendix to: Tests of Equal Forecast Accuracy and Encompassing for Nested Models', mimeo.
- [7] Clark, T.E. and M. W. McCracken (2001), 'Tests of Equal Forecast Accuracy and Encompassing for Nested Models', *Journal of Econometrics*, vol.105, 85-110.

- [8] Clark, T.E. and M. W. McCracken (2005), ‘Evaluating Direct Multistep Forecasts’, *Econometric Reviews*, vol.24, 369-404.
- [9] Clark, T.E. and M. W. McCracken (2011), ‘Reality Checks and Nested Forecast Model Comparisons’, *Journal of Business & Economic Statistics*, forthcoming.
- [10] Clark, T.E. and M. W. McCracken (2011b), ‘Advances in Forecast Evaluation’, manuscript, St. Louis Federal Reserve Bank.
- [11] Clark, T.E. and K. West (2006), ‘Using out-of-sample Mean Squared Prediction Errors to test the Martingale Difference Hypothesis’, *Journal of Econometrics*, vol.138, 291-311.
- [12] Clark, T.E. and K. West (2007), ‘Approximately Normal Tests for Equal Predictive Accuracy in Nested Models’, *Journal of Econometrics*, vol.138, 291-311.
- [13] Cogley, T. and T.J. Sargent (2001), ‘Evolving Post-World War II U.S. Inflation Dynamics’, *NBER Macroeconomic Annual*.
- [14] Cooper, M. and H. Gulen (2006), ‘Is Time-Series-Based Predictability Evident in Real Time?’ *The Journal of Business*, vol. 79, 1263-1292.
- [15] Corradi, V. and N. R. Swanson (2002), ‘A Consistent Test for Nonlinear Out-of-Sample Predictive Accuracy’, *Journal Econometrics*, vol.110, 353-381.
- [16] Corradi, V. and N. R. Swanson (2007), ‘Nonparametric Bootstrap Procedures for Predictive Inference based on Recursive Estimation Schemes’, *International Economic Review* vol.48, 67-109.
- [17] Diebold, F.X., and R.S. Mariano (1995), ‘Comparing Predictive Accuracy’, *Journal of Business and Economic Statistics*, vol.13, 253-263.
- [18] Giacomini R. and H. White (2006), ‘Tests of Conditional Predictive Ability’, *Econometrica*, vol. 74, 1545-1578.
- [19] Goncalvez, S. and L. Kilian (2004), ‘Bootstrapping autoregressions with conditional heteroskedasticity of unknown form’, *Journal of Econometrics*, 123, 89-120.
- [20] Goyal A. and I. Welch (2008), ‘A Comprehensive Look at the Empirical Performance of Equity Premium Prediction’, *Review of Financial Studies*, vol. 21, 1455-1508.

- [21] Guo, H. (2006), ‘On the Out-of-sample Predictability of Stock Market Returns’, *The Journal of Business*, vol. 79, 645-670.
- [22] Harvey, D.I., S.J. Leybourne and P. Newbold (1998), ‘Tests for Forecast Encompassing’, *Journal of Business and Economic Statistics*, vol.16, 254-259.
- [23] Hendry, D. F. and K.Hubrich (2011), ‘Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate’, *Journal of Business and Economic Statistics*, vol.29 (2), 216-227.
- [24] Hubrich, K. (2005), ‘Forecasting Euro Area Inflation: Does Aggregating Forecasts by HICP Component Improve Forecast Accuracy?’, *International Journal of Forecasting*, 21(1), 119-136, 2005.
- [25] Hubrich, K. and K. D. West (2010), ‘Forecast Evaluation of Small Nested Model Sets’, *Journal of Applied Econometrics*, vol.25, 574-594.
- [26] McCracken, M. W. (2007), ‘Asymptotics for Out of Sample Tests of Causality’, *Journal of Econometrics*, vol. 140, 719-752.
- [27] Meese, R.A. and K. Rogoff (1983), ‘Empirical Exchange Rate Models of the Seventies: do they fit out of sample?’, *Journal of International Economics*, vol.14, 3-24.
- [28] Newey, W. K., West, K. D. (1987), A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, vol 55, 703-708.
- [29] Ravazzolo, F. and P. Rothman (2010), ‘Oil and US GDP: A Real-Time Out-of-Sample Examination’, mimeo.
- [30] Silvapulle M. J. and P. K. Sen (2005), ‘Constrained Statistical Inference: Inequality, Order, and Shape Restrictions’, Wiley-Interscience.
- [31] Stock, J. H., and M. W. Watson (1999), ‘Forecasting Inflation’, *Journal of Monetary Economics*, vol. 22, 293-335.
- [32] Stock, J. H., and M. W. Watson (2003), ‘Forecasting Output and Inflation: the Role of Asset Prices’, *Journal of Economic Literature*, vol.41, 788-829.
- [33] Stock, J. H., and M. W. Watson (2010), ‘Modeling Inflation After the Crisis’, mimeo.

- [34] West, K.D. (1996), 'Asymptotic Inference about Predictive Ability', *Econometrica*, vol.64, 1067-1084.
- [35] West, K.D. (2006), 'Forecast Evaluation', in *Handbook of Economic Forecasting*, vol.1, 100-134, Elsevier.
- [36] White, H. (2000), 'A Reality Check for Data Snooping', *Econometrica*, vol.68, 1097-1126.