Granziera, Eleonora; Hubrich, Kirstin; Moon, Hyungsik Roger

Working Paper

A predictability test for a small number of nested models

ECB Working Paper, No. 1580

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Granziera, Eleonora; Hubrich, Kirstin; Moon, Hyungsik Roger (2013) : A predictability test for a small number of nested models, ECB Working Paper, No. 1580, European Central Bank (ECB), Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/154013

EUROPEAN CENTRAL BANK

EUROSYSTEM

In 2013 all ECB publications feature a motif taken from the €5 banknote.

WORKING PAPER SERIES

NO 1580 / AUGUST 2013

A PREDICTABILITY TEST FOR A SMALL NUMBER OF NESTED MODELS

Eleonora Granziera, Kirstin Hubrich and Hyungsik Roger Moon

Acknowledgements
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, EUI Conference and two anonymous referees for helpful comments and suggestions. The views expressed are those of the authors and do not represent those of the BoC or ECB.

Eleonora Granziera
Bank of Canada; e-mail: egranziera@bankofcanada.ca

Kirstin Hubrich
European Central Bank; e-mail: kirstin.hubrich@ecb.europa.eu

Hyungsik Roger Moon
University of Southern California; e-mail: moonr@usc.edu

Address	Kaiserstrasse 29, 60311 Frankfurt am Main, Germany
Postal address	Postfach 16 03 19, 60066 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Internet	http://www.ecb.europa.eu
Fax	+49 69 1344 6000

ISSN	1725-2806 (online)
EU Catalogue No	QB-AR-13-077-EN-N (online)

Abstract

In this paper we introduce Quasi Likelihood Ratio tests for one sided multivariate hypotheses 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 existing equal and superior predictive ability tests for multiple model comparison. We find that our proposed tests are well sized for one step ahead as well as for multi-step ahead forecasts when critical values are bootstrapped. The experiments on the power reveal that the superior predictive ability test performs last while the ranking between the quasi likelihood-ratio test and the other equal predictive ability tests 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.

Non-technical Summary

Given the forward looking nature of monetary policy, forecasting future economic variables is central for policy makers. Developing forecasting models to obtain predictions for variables of interest such as inflation and GDP growth is key for both monetary policy decision makers and policy observers. Often, researchers consider a range of competing forecasting models using different methods or emphasizing different aspects of the economy, say, labour markets or financial markets. Once forecasts from these models are produced it is important to assess their accuracy. Then, evaluation of forecast accuracy requires comparing the performance of a *set* of models. The criterion often used for the comparison is the mean squared prediction error (MSPE), which measures the difference between the predictions and the realizations of the forecasted variables. Testing whether the models provide the same forecast accuracy, i.e. the same MSPE, represents a test of equal predictive ability.

Two important distinctions are relevant for the contribution of this paper: First, most of the previous literature focusses on pairwise forecast model comparison. Instead, this paper focusses on comparing forecast accuracy of multiple models simultaneously. The contribution of this paper lies in 1) suggesting a novel equal predictive ability test in this context, and 2) providing a review of the few contributions to the literature of multiple forecast model comparison in a unified notational framework and comparing their finite sample properties via an extensive Monte Carlo simulation exercise. In multiple model comparisons the models are simultaneously evaluated against one particular model in the set. This model, chosen by the researcher, is called the benchmark and is usually the most parsimonious. A second important distinction in the literature on forecast accuracy testing is whether the alternative models nest the benchmark or not. An alternative forecast model nests the benchmark if, for example, it contains at least all the predictors of the benchmark model. The distinction between nested and non-nested forecast model comparisons is important for choosing the appropriate test procedure for testing equal predictive ability. While we review tests for multiple forecast model comparisons for models that nest and for models that do not nest the benchmark model, our newly suggested test statistic is appropriate for the comparison of forecasts from a set of alternative models that nest the benchmark. A novel feature is that in formulating the alternative hypothesis and the test statistics we distinguish among three cases in the nesting structure of the alternative models.

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. In the existing literature Hubrich and West (2010) (hereafter HW) consider this setup and propose two approaches: one is to directly extend the pairwise model comparison in Diebold and Mariano (1995) and West (1996) to a chi-squared statistic, and the other is to take the maximum of t-statistics (max-t test) of all the pairwise MSPE differences adjusted for estimation uncertainty. Also for nested multiple model comparison Inoue and Kilian (2005) derive the asymptotic distribution of two tests of predictability for one step ahead forecasts. Clark and McCracken (2012), thereafter CM, suggest two additional tests and a new bootstrap procedure to approximate the asymptotically valid critical values of the new and existing tests for multiple model comparison of nested predictive models.

The main contribution of the paper is to propose an alternative test to the ones in HW and CM. We propose a new one-sided quasi-likelihood ratio (hereafter QLR) predictability test for the comparison of a small number of models nesting the benchmark model. The QLR test

statistic depends on the structure of the alternative models. We distinguish among three different cases: (i) when the alternative 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. This distinction is aimed at improving the power of the test. We derive the asymptotic distribution of the tests and find that it is non-standard and depends 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 we consider two approaches, (i) bootstrapping and (ii) simulations based on the normal approximation of the MSPE difference estimates. We prove the validity of the bootstrap procedure developed in CM for our proposed test. As a second contribution of our paper we discuss the tests of equal and superior predictive accuracy for multiple model comparisons suggested in the literature in a unified notational framework.

The finite sample size and power properties of the tests are evaluated via extensive Monte Carlo simulations for one and four-step ahead forecasts. We find that our proposed tests are well sized for one step ahead as well as for multi-step ahead forecasts when critical values are bootstrapped. Using the simulated normal critical values provides reasonable size for one-step ahead and for the maximum t-statistic also for four-step ahead. The experiments on the power reveal that the superior predictive ability test performs last while the ranking between the quasi likelihood-ratio test and the other equal predictive ability tests depends on the simulation settings.

Finally, we present an empirical analysis where we find that the recessionary gap and the food and energy inflation 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.

1 Introduction

Evaluation of forecast accuracy usually requires comparing 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 non-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; Hansen (2005) suggests standardizing the White (2000) statistic to achieve better power. 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 (2006), 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 condition[1] (e.g., 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 tests 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, 2005a) 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 well sized tests even when critical values are obtained under the

[1] Under 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.

normal approximation. They also provide Monte Carlo evidence supporting their suggested procedure.

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. In the existing literature Hubrich and West (2010) (hereafter HW) consider this setup 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. Again for nested multiple model comparison Inoue and Kilian (2005) derive the asymptotic distribution of two tests of predictability for one step ahead forecasts. Clark and McCracken (2012), thereafter CM, suggest a new bootstrap procedure to approximate the asymptotically valid critical values for two tests of equal MSPE and two tests of forecast encompassing for multiple model comparison of nested predictive models.

The main contribution of the paper is to propose an alternative test to the ones in HW and CM. 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. By treating the MSPE differences as a multivariate parameter we formulate the problem of testing for equal predictability as testing a multivariate parameter that takes one-sided values. Then, we propose one-sided quasi-likelihood ratio (hereafter QLR) predictability tests for the comparison of a small number of models nesting the benchmark model. The QLR test statistic depends on the structure of the alternative models. We distinguish among three different cases: (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 we consider two approaches, (i) bootstrapping and (ii) simulations based on the normal approximation of the MSPE difference estimates.

As a second contribution of our paper we discuss the tests of equal and superior predictive accuracy for multiple model comparisons suggested in the literature in a unified notational

framework. 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 the QLR type tests, the max-t test of HW and the tests in CM. We also include the tests of superior predictive ability by White (2000), Hansen (2005) and Corradi and Swanson (2007). Our Monte Carlo investigation reveals that, for critical values derived through the fixed regressor bootstrap, as in CM, the QLR and max-t tests are correctly sized for one step ahead forecasts and the size distortions for the longer forecast horizons are generally not severe. Also, as previously found by HW, the use of simulated critical values based on the normal approximation of the MSPE differences overall performs well in terms of size in the case of one-step ahead forecasts when the test statistics are adjusted as in CW and HW. For four step ahead forecasts and for small out-of-sample forecast evaluation periods, the QLR tests are oversized, however, and for short out-of-sample periods even grossly oversized, while the max-t statistic is only somewhat oversized for short out-of-sample periods, and otherwise exhibits reasonable size properties. The White, Hansen and Corradi and Swanson superior predictive ability tests perform poorly in terms of size as they are not suited for nested model comparison. As far as the power of the test is concerned, the ranking between the quasi likelihood-ratio 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.

As an illustrative application, we evaluate equal predictive ability for forecasting the US CPI core yearly inflation rate for an AR(1) model as a benchmark and three other alternative models that extend the benchmark 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. Section 3 presents the tests. Section 4 provides procedures for inference based on the tests. In Section 5 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 6. Section 7 concludes.

2 Testing Framework

2.1 Notation and General Setup

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

Suppose that one is interested in forecasting a scalar $y_{t+\tau}$, $\tau \geq 1$, using $M + 1$ linear[2] models:

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

where $x'_{m,t}\beta_m$ is the linear projection of $y_{t+\tau}$ on the predictor $x_{m,t}$ and $u_{m,t+\tau}$ denotes the population forecast error with zero mean[3] and satisfying $E(u_{m,t+\tau}x_{m,t}) = 0$ for $m = 0, 1, ..., M$. Note 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.

Denote by $\hat{y}_{0,t+\tau}, ..., \hat{y}_{m,t+\tau}, ..., \hat{y}_{M,t+\tau}$ the $\tau-$period ahead forecasts obtained from the estimated models either through the expanding window or the rolling scheme[4] for $t = T, ..., T + P - \tau$. Here $T + P$ is the total sample size, 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 $x_{0,t}, ..., x_{M,t}$ are vectors of predictors such that $x_{0,t} = x_{01,t}$ is of dimension $k_0 \times 1$ and $x_{m,t} = \left(x'_{01,t},\ x'_{m2,t}\right)'$ is of dimension $k_m \times 1$ with $k_0 < k_m$.

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

[3] It is implicitly assumed that all models include an intercept term.

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

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, $m \in \{1, ..., M\}$. Under the null hypothesis, model 0 is the "true" model in the sense that each model m includes $k_m - k_0$ excess parameters:

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

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

$$f_{t+\tau} = \left(f_{1,t+\tau}, ..., f_{M,t+\tau}\right)'.$$

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

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

with $\sigma_m^2 \equiv E\left(u_{m,t+\tau}^2\right)$ 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 by:

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

and

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

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

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

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 $(P-\tau+1)^{-1}\sum_{t=T}^{T+P-\tau}\left(\hat{f}_{m,t+\tau}\right) < 0$ in finite samples. To improve the finite sample properties, they suggest using an adjusted sample MSPE to center it around zero, as

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

where $\left(\hat{y}_{0,t+\tau} - \hat{y}_{m,t+\tau}\right)^2$ is the adjustment term. In the Appendix we show 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). This implies that our proposed QLR test statistics based on adjusted MSPE can be interpreted as encompassing tests for small nested model sets. Analogous quantities defined above for $\hat{f}_{m,t+\tau}$ can be derived from $\hat{f}_{m,t+\tau}^{adj}$:

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

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.[5]

2.2 Hypotheses

Given the parameterization of predictability with the average difference of the SPE in (2), it is natural to express the null hypothesis that the parsimonious model, model 0, performs

[5]Note that Clark and McCracken (2005b) show analytically and in simulations that the adjusted and unadjusted statistics have different behaviour in the presence of unmodeled structural change under the alternative.

equally well as a larger model, say model m, $m \in \{1, ..., M\}$ as

$$H_0 : \mu = 0,$$

or equivalently

$$H_0 : \mu^{adj} = 0.$$

The specification of the alternative hypothesis will depend on the assumptions about the nesting structure of the alternative models. In this paper, 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) a general case in which the models are nested within each group but not across groups.

2.2.1 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: $x_{m,t} = \left(x'_{m-1,t},\ X'_{m,t}\right)'$ so that $k_0 < \cdots < k_m < \cdots < k_M$.

Given the structure of the problem, if model m^* is the true model, then for models $m = 1, ..., m^* - 1$, it will hold that $\sigma^2_{m^*} < \sigma^2_{m^*-1} \leq \cdots \leq \sigma^2_1$ and hence $0 \leq \mu_1 = \sigma^2_0 - \sigma^2_1 \leq \cdots \leq \sigma^2_0 - \sigma^2_{m^*-1} = \mu_{m^*-1} < \sigma^2_0 - \sigma^2_{m^*} = \mu_{m^*}$, while for models $m^* + 1, ..., M$, it will be the case that $\sigma^2_{m^*} = \sigma^2_{m^*+1} = ... = \sigma^2_M$, which implies $\mu_{m^*} = \mu_{m^*+1} = ... = \mu_M$. Note this holds only for the case in which the set of regressors is progressively expanding with the models,[6] 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 alternative hypotheses can be expressed as:[7]

$$H_1 : 0 \leq \mu_1 \leq \mu_2 \leq \cdots \leq \mu_M,\ \mu \neq 0, \tag{3}$$

or equivalently as

$$H_1 : D\mu \geq 0,\ \mu \neq 0, \tag{4}$$

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

[7] Equivalently, since this ordering is invariant to the introduction of the CW adjustment the alternative can be expressed with respect to μ^{adj} for each of the three cases considered: (i) $H_1 : 0 \leq \mu^{adj}_1 \leq \mu^{adj}_2 \leq \cdots \leq \mu^{adj}_M$, $\mu^{adj} \neq 0$; (ii) $H_1 : \mu^{adj} \geq 0, \mu^{adj} \neq 0$; (iii) $H_1 : D^B\mu^{adj} \geq 0$, $\mu^{adj} \neq 0$.

where

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ & & \ddots & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \tag{5}$$

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).

2.2.2 Non-Nested Alternative Models

In this case there is no nesting relation between the alternative models, but still each of them nests the benchmark. Then, the alternative hypothesis can be expressed as

$$H_1 : \mu_1 \geq 0, ..., \ and \ \mu_M \geq 0, \ \mu \neq 0, \tag{6}$$

or equivalently as

$$H_1 : \mu \geq 0, \ \mu \neq 0. \tag{7}$$

2.2.3 General Case: 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: $\mu_{k,1} \leq \mu_{k,2} \leq \cdots \leq \mu_{k,M_k}$, with M_k the number of models included in group k. Here groups can have common alternative models.[8] Define $\mu^k = \left(\mu_{k,1}, \mu_{k,2}, \cdots, \mu_{k,M_k}\right)$. Then, the alternative hypothesis is expressed as

$$H_1 : D_1\mu^1 \geq 0, ..., \ and \ D_K\mu^K \geq 0, \ \left(\mu^{1\prime}, ..., \mu^{K\prime}\right)' \neq 0,$$

or equivalently as

$$H_1 : D^G\mu \geq 0, \ \mu \neq 0. \tag{8}$$

[8] For example, when $M = 3$, if models 1 and 2 are nested by model 3 but models 1 and 2 do not nest each other, then $K = 2$, where the first group consists of models 1 and 3 and the second group consists of models 2 and 3. In this case, model 3 is shared by the two groups.

for some matrix D^G whose entries are one of $-1, 0$, and 1.[9] In a special case where each group is mutually exclusive, the alternative hypothesis becomes

$$H_1 : D^B \mu \geq 0, \ \mu \neq 0,$$

where

$$D^B = diag\left(D_1, ..., D_K\right).^{10}$$

with D_k a $M_k \times M_k$ matrix defined as D in (5).

3 Test Statistics

3.1 Quasi-Likelihood Ratio (QLR) Test Statistic

In the three cases of the previous subsections, we can express the null and the alternative hypothesis in a general form as

$$H_0 : \mu = 0 \text{ (or } \mu^{adj} = 0)$$

vs

$$H_1 : G\mu \geq 0 \text{ and } \mu \neq 0 \ \left(\text{or } G\mu^{adj} \geq 0 \text{ and } \mu^{adj} \neq 0\right) \tag{9}$$

for some matrix G. When there are multiple restrictions in G (that is, the number of the rows of G is larger than one), the alternative hypothesis in (9) is a multivariate one-sided hypothesis.

Let

$$\mathcal{A}_0 = \{0\} \text{ and } \mathcal{A} = \{\mu : G\mu \geq 0\} \ \left(\text{or } \mathcal{A} = \left\{\mu^{adj} : G\mu^{adj} \geq 0\right\}\right).$$

Here $\mathcal{A}_0$ is the parameter set under the null and $\mathcal{A}$ is the maintained parameter set (the union of the null parameter set and the alternative parameter set), and we can reexpress the null and the alternative hypotheses as

$$H_0 : \mu \in \mathcal{A}_0 \text{ vs } H_1 : \mu \in \mathcal{A} - \mathcal{A}_0. \tag{10}$$

[9] For example, in the case of the footnote above, $D^G = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$.

[10] Matrix $diag\left(D_1, ..., D_K\right)$ is the block diagonal matrix whose diagonal blocks are $D_1, ..., D_K$.

Notice that the maintained parameter set $\mathcal{A}$ is a convex cone.

There is a long history in statistics literature that studied testing for the hypotheses expressed in (10). For example,[11] Perlman (1969) studied the likelihood ratio test when μ is the mean of iid multivariate Gaussian random vectors with unknown variance and the parameter sets $\mathcal{A}_0$ and $\mathcal{A}$ are positively homogenous.

Following Perlman (1969)'s idea, to test for (9) or equivalently (10), we propose a quasi-likelihood ratio statistic that is based on the standard Gaussian likelihood ratio statistic when $\bar{f}^{adj} \sim N\left(\mu, \frac{1}{P-\tau+1}V\right)$:

$$\begin{aligned} & QLR \\ = \; & (P-\tau+1)\min_{\mu\in\mathcal{A}_0}\left(\bar{f}^{adj}-\mu\right)'\hat{W}\left(\bar{f}^{adj}-\mu\right) - (P-\tau+1)\min_{\mu\in\mathcal{A}}\left(\bar{f}^{adj}-\mu\right)'\hat{W}\left(\bar{f}^{adj}-\mu\right) \\ = \; & (P-\tau+1)\,\bar{f}^{adj\prime}\hat{W}\bar{f}^{adj} - (P-\tau+1)\min_{G\mu\geq 0}\left(\bar{f}^{adj}-\mu\right)'\hat{W}\left(\bar{f}^{adj}-\mu\right), \end{aligned}$$

where $\hat{W}$ is a general weighting matrix whose limit (either the weak or the probability limit) is strictly positive definite (a.s.). Examples of widely used weight matrices are $\hat{W}=\hat{V}^{-1}$ or $\hat{W}=diag\left(\hat{v}_1,...,\hat{v}_M\right)^{-1}$, where as an estimator $\hat{V}$, we use the Newey-West (1987) HAC estimator

$$\hat{V}=\hat{\Gamma}_0+\sum_{j=1}^{\tau-1}\left[1-\frac{j}{(\tau-1)+1}\right]\left(\hat{\Gamma}_j+\hat{\Gamma}_j'\right). \tag{11}$$

and $\hat{\Gamma}_i=\frac{1}{P-\tau+1}\sum_{t=T}^{T+P-\tau}\left(\hat{f}_t^{adj}-\bar{f}^{adj}\right)\left(\hat{f}_{t+i}^{adj}-\bar{f}^{adj}\right)'$, and $\hat{v}_m$ denotes the m^{th} diagonal element of $\hat{V}$.

We call it a quasi-likelihood ratio statistic (rather than a likelihood ratio statistic) because it is based on misspecified likelihood function (the true distribution of $\bar{f}^{adj}$ is not $N\left(\mu, \frac{1}{P-\tau+1}V\right)$). It follows the same spirit of the quasi maximum likelihood estimator of White (1982, 1994).

Throughout this paper, notation QLR_D, QLR_I, and QLR_{D^G} denotes QLR with $G = D, I, D^G$, respectively to differentiate among the structure of the alternative models.

3.2 Alternative Nested Models Tests

We consider few alternative forecast accuracy tests for nested multi-model comparison considered in the existing literature proposed by HW and CM.

[11] Sillvapulle and Sen (2005) and Andrews (2001) are examples of more recent studies on this problem.

HW suggest the test statistics:

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

$$max-t-unadj=\max_{1\leq m\leq M}\left\{\sqrt{(P-\tau+1)}\frac{\bar{f}_1}{\sqrt{\hat{v}_1}},...,\sqrt{(P-\tau+1)}\frac{\bar{f}_M}{\sqrt{\hat{v}_M}}\right\},$$

which is the maximum of the t-statistics where $\hat{v}_m$ is the m^{th} diagonal element of $\hat{V}$ in (11) .

CM consider the additional test statistics:

$$max-F=\max_{1\leq m\leq M}\left\{(P-\tau+1)\frac{\bar{f}_1^{adj}}{\hat{\sigma}_1^2},...,(P-\tau+1)\frac{\bar{f}_M^{adj}}{\hat{\sigma}_M^2}\right\},$$

$$max-F-unadj=\max_{1\leq m\leq M}\left\{(P-\tau+1)\frac{\bar{f}_1}{\hat{\sigma}_1^2},...,(P-\tau+1)\frac{\bar{f}_M}{\hat{\sigma}_M^2}\right\},$$

where $\bar{f}_m$ denotes the $m-th$ element in the vector $\bar{f}$ and $\hat{\sigma}_m^2=\sum_{t=T}^{T+P-\tau}\frac{1}{P-\tau+1}\hat{u}_{m,t+\tau}^2$. These last two statistics were also considered in Inoue and Kilian (2005) who derive the asymptotic distribution of max-F and max-F-unadj for the case of one step ahead predictions.

3.3 Superior Predictive Ability Tests

Although the hypothesis of interest is different, for completeness we consider three additional tests for multiple model comparison developed by White (2000), Hansen (2005) and Corradi and Swanson (2007). These tests involve the composite null hypothesis:

$$H_0:\mu\leq 0. \tag{12}$$

Because the null states that the benchmark is superior or equal (not inferior) to the alternative models they are called superior predictive ability tests. Given the null hypotesis in (12), these tests are not designed for nested model comparison but rather accommodate cases in which at least one of the alternative models is non-nested with the benchmark model.

White (2000) uses the test statistic:

$$SPA_W=\max_{1\leq m\leq M}\left\{\sqrt{(P-\tau+1)}\bar{f}_1,...,\sqrt{(P-\tau+1)}\bar{f}_M\right\} \tag{13}$$

where $\bar{f}_m$ is the sample mean SPE associated with model m.

The other two tests are recent variants of the White (2000) reality check test: Hansen (2005) suggests the standardized test statistic:

$$SPA_H = \max\left[\max_{1\leq m\leq M}\left\{\sqrt{(P-\tau+1)}\frac{\bar{f}_1}{\sqrt{\hat{v}_1}},...,\sqrt{(P-\tau+1)}\frac{\bar{f}_M}{\sqrt{\hat{v}_M}}\right\},0\right],$$

where $\hat{v}_m$ is a consistent estimator of the $m-th$ diagonal element of $\hat{V}$. This test statistic is designed to discard poor and irrelevant alternatives from the set of forecasting models. As in White (2000) the critical values are obtained through a stationary bootstrap.

Corradi and Swanson (2007) generalize White (2000) to the case in which the effect of parameter estimation error does not vanish asymptotically. Under a quadratic loss function they formulate their test statistic as in (13), but they propose a new nonparametric block bootstrap procedure to account for model misspecification and non-vanishing parameter estimation error.

4 Limiting Distribution and 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, and this makes it difficult to use the critical values of the limiting 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 QLR statistic since it covers a general case. It is straightforward to modify these approaches for the test statistics QLR_D, QLR_I and QLR_{D^G}.

4.1 Limiting Distribution of QLR Statistic

We derive the asymptotic distribution of the general QLR test statistic

$$QLR = (P-\tau+1)\,\bar{f}^{adj\prime}\hat{W}\bar{f}^{adj} - (P-\tau+1)\min_{G\mu\geq 0}\left(\bar{f}^{adj}-\mu\right)'\hat{W}\left(\bar{f}^{adj}-\mu\right).$$

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 $m = 0, 1, ..., M$. Define $u_{t+\tau} = y_{t+\tau} - x_t'\beta_0$, where $x_t'\beta_0$ is the (population) projection of $y_{t+\tau}$ on x_t. Then, under the null hypothesis, $u_{0,t} = u_{1,t} = \cdots = u_{M,t} = u_t$. Let $u_t = u_{m,t}$ for all t and $h_{t+\tau} = x_t u_{t+\tau}$. Denote $H_t = \frac{1}{t}\sum_{s=1}^{t-\tau} h_{t+\tau}$, $\Sigma_x = E(x_t x_t')$, and $\Omega_h = \lim_T \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E(h_t h_s')$. Define $\tilde{h}_t = \Omega_h^{-1/2} h_t$ and $Q_m = \Omega_h^{1/2}\left(J_m'\left(J_m\Sigma_x J_m'\right)^{-1} J_m - J_0'\left(J_0\Sigma_x J_0'\right)^{-1} J_0\right)\Omega_h^{1/2}$. We use $W(r)$ to denote the k_x dimensional Wiener process.

We make the following assumptions which are quite standard in the literature (e.g., Clark and McCracken (2001, 2005a, 2011), McCracken (2007)).[12]

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

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

Denote $U_{t+\tau} = \left(h_{t+\tau}', vech\left(x_t x_t' - E\left(x_t x_t'\right)\right)\right)'$.

Assumption 2 *(a) U_t is strictly stationary with $E(U_t) = 0$ and $E\|U_t\|^r < \infty$, for some $r > 8$. (b) $E\left(h_t h_{t-j}'\right) = 0$ for all $j \geq \tau$. (c) $E\left(x_t x_t'\right) > 0$. (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.*

Assumption 3 *Assume that* $\lim_{P,T\to\infty} \frac{P}{T} = \lambda \in (0, \infty)$.

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

$$\mu(\alpha, \Psi) = \arg\min_{G\mu \geq 0} (\alpha - \mu)' \Psi (\alpha - \mu).$$

Then, it is well known that

$$\alpha'\Psi\alpha - \min_{G\mu\geq 0} (\alpha - \mu)' \Psi (\alpha - \mu) = \mu(\alpha, \Psi)' \Psi \mu(\alpha, \Psi)$$

(e.g., Silvapulle and Sen (2005) , Proposition 3.4.1) and $\mu(\alpha, \Psi)$ is continuous in (α, Ψ). (e.g., see page 213 of Silvapulle and Sen(2005)).

Define

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

[12] Refer to these studies for a detailed discussion of the assumptions.

Theorem 1 *Assume Assumptions 1 – 3. Then, under the null hypothesis, we have*

$$\frac{1}{2}\left(P-\tau+1\right)\bar{f}^{adj}\Rightarrow\mathcal{F}.$$

Furthermore, if

$$\left[\begin{array}{c}\frac{1}{2}\left(P-\tau+1\right)\bar{f}^{adj}\\ \hat{W}\end{array}\right]\Rightarrow\left[\begin{array}{c}\mathcal{F}\\ \mathcal{W}\end{array}\right], \tag{14}$$

where $\mathcal{W}$ is strictly positive definite with probablity one, then,

$$QLR\Rightarrow\mu\left(\mathcal{F},\mathcal{W}\right)'\mathcal{W}\mu\left(\mathcal{F},\mathcal{W}\right).$$

Remarks:

1. In the appendix we show that under the assumptions in Theorem 1, if the null is true, then

   $$\hat{V}\Rightarrow\mathcal{V},$$

   where

   $$\mathcal{V}=\left[\begin{array}{ccc}\mathcal{V}_{11} & \cdots & \mathcal{V}_{1M}\\ \vdots & \ddots & \vdots\\ \mathcal{V}_{M1} & \cdots & \mathcal{V}_{MM}\end{array}\right],$$

   and

   $$\mathcal{V}_{mn}=\int_{1}^{1+\lambda}\frac{1}{r^{2}}W\left(r\right)'Q_{m}Q_{n}W\left(r\right)dr.$$

2. Furthermore, suppose that B is an $(M\times M)$ matrix such that its $(m,n)^{th}$ element is $\mathrm{tr}\left(Q_{m}Q_{n}\right)$, that is,[13] $B=\left[\mathrm{tr}\left(Q_{m}Q_{n}\right)\right]_{(m,n)}$. If rank $rank\left(B\right)=M$, then $\mathcal{V}$ is strictly positive definite a.s.

3. If $\mathcal{W}$ is a positive definite non-random matrix, then the joint limit condition (14) is implied by $\hat{W}\rightarrow_{p}\mathcal{W}>0$.

Note that the limiting distribution of QLR is a functional of Brownian motion and it depends on the characteristics of the data generating process such as the out-of-sample to

[13] Note that in a special case where $\Omega_{h}=\Sigma_{x}=I_{k_{x}}$, $\mathrm{tr}\left(Q_{m}Q_{m}\right)$ is $k_{m}-k_{0}>0$ and $\mathrm{tr}\left(Q_{m}Q_{n}\right)=$ (# of the common regressors in models m and n) $-\,k_{0}$. In this case, if the alternative models are non-nested, then B is a diagonal matrix with positive diagonal elements.

in-sample ratio and on the covariance matrix of the regressors that do not overlap. Therefore critical values should be tabulated for every application.[14]

4.2 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 (2012) 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 k_x predictors. Then, compute the residuals $\hat{u}_{s+\tau} = y_{s+\tau} - \hat{\beta}' x_{s+\tau}$, for $s = 1, ..., T+P-\tau$.

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

Step 3: Estimate the benchmark 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 $s = 1, \cdots, T+P-\tau$.

Step 4: Using $\left\{y^*_{s+\tau}, x_s\right\}_{s=1,...,T+P-\tau}$, construct the test statistic QLR^*.

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

Consider the following assumption:[15]

Assumption 4 *(a) Under the null, the forecast error u_t is an invertible $MA(\tau-1)$ process generated by $u_t = \varepsilon_t + \theta_1^0 \varepsilon_{t-1} + \cdots + \theta_{\tau-1}^0 \varepsilon_{t-\tau+1}$, where $\varepsilon_t \sim iid$ with $E(\varepsilon_t) = 0$, $E\|\varepsilon_t\|^r < \infty$, for some $r > 8$, and $\varepsilon_0 = \cdots = \varepsilon_{1-\tau} = 0$. (b) Denote $\Theta(L;\theta) = 1 + \theta_1 L + \cdots + \theta_{\tau-1} L^{\tau-1}$. 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 $\left(\theta^0, \beta^0\right)$ 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.*

[14] In the appendix available from www-rcf.usc.edu/~moonr, we provide a procedure to simulate the asymptotic critical values and simulation results regarding the small sample properties of the QLR tests evaluated against the asymptotic critical values.

[15] Assumption 4(a) implies that the forecasts are optimal as it requires the τ step ahead forecast errors to be at most MA($\tau-1$) processes.

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

Theorem 2 *Assume Assumptions 1 – 4. Then,* $\frac{1}{2}(P-\tau+1)\,\bar{f}^{*adj} \Rightarrow^* \mathcal{F}$. *Furthermore, if* $\left(\frac{1}{2}(P-\tau+1)\,\bar{f}^{*adj}, \hat{W}^*\right) \Rightarrow^* (\mathcal{F}, \mathcal{W})$, *where* $\mathcal{W}$ *is positive definite a.s., then we have*

$$QLR^* \Rightarrow^* \mu(\mathcal{F}, \mathcal{W})' \mathcal{W} \mu(\mathcal{F}, \mathcal{W}).$$

4.3 Use of Normal Approximation

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

However, under different set-ups,[16] it is possible to approximate $\bar{f}^{adj}_m/\sqrt{\hat{v}_m}$ with a normal distribution (see Giacomini and White (2006), Clark and McCracken (2001, 2005a), and the discussion in HW). Using the normal approximation of the distribution of $\bar{f}^{adj}_m/\sqrt{\hat{v}_m}$, HW proposed an inference based on the maximum of correlated normals, building on results from the literature of order statistics. For one-step ahead forecasts and homoscedastic prediction errors the simulation experiments in CW and HW provide evidence that the size properties of their test statistics with the standard normal approximation of $\bar{f}^{adj}_m/\sqrt{\hat{v}_m}$ are reasonable. They also demonstrate that the standard normal approximation performs reasonably well in heteroscedastic environments when the number of additional regressors, k_m is equal to one. 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 of CW and HW, one may conjecture that treating $\bar{f}^{adj}$ as normal might deliver a reasonable approximation despite the limiting distribution of the quasi likelihood ratio test being a functional of Brownian motion under the maintained assump-

[16] Example of these are when the null hypothesis is to test the forecasting methods not models and when one uses an asymptotic approximation holding T fixed and letting $P \to \infty$ (Giacomini and White (2006)), and when $\lambda = 0$ (Clark and McCracken (2001, 2005)).

tions of the paper.[17] This leads us to use the critical values computed under the normality assumption for the MSPE-adjusted to evaluate the tests in our simulations. These critical values can be computed as follows. Treat $\bar{f}^{adj} \sim N(0, V)$. Define $Z = (Z_1, ..., Z_M)' \sim N\left(0, \hat{R}^{-1/2}\hat{V}\hat{R}^{-1/2}\right)$, where $\hat{R} = diag(\hat{V})$. Then, we approximate the distributions of the tests as follows: $\max_{1\leq m\leq M}\{Z_1, ..., Z_M\}$ for the limit of $max-t$, and $\mu\left(Z, \hat{W}\right)'\hat{W}\mu\left(Z, \hat{W}\right)$ for the limit of QLR. Though this approach might be appealing because it is easy to implement, we stress that under the maintained assumptions of this paper the limiting null distribution of the MSPE differences is non-normal and therefore the Normal approximation is not guaranteed to deliver well-sized tests.

5 Monte Carlo Simulation

In this section we first 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 critical values derived through simulations by bootstrapping or by assuming normality of the MSPE. We present results for test statistics based on both unadjusted and adjusted MSPEs. Additionally, we provide results for the superior predictive ability tests discussed in Section 3.3.

In this simulation study we go beyond the existing literature in presenting new small sample evidence for our proposed QLR tests using different critical values, including normal and bootstrapped critical values. We add to the simulation evidence in HW and CM results for different DGPs. In addition, we present new evidence for the superior predictive ability test suggested by Corradi and Swanson (2007), SPA_CS, that to our knowledge has not been presented in simulation studies so far.

5.1 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.

[17] We stress that these critical values might be incorrect under our approximation framework, so we do not claim validity of the normal approximation for our test statistics.

The design for DGP1 takes the form:

$$y_{t+\tau} = c + \rho y_t + \gamma' x_t + u_{t+\tau} \tag{15}$$

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:[18] $u_{t+\tau} = \varepsilon_{t+\tau} + 0.95\varepsilon_{t+\tau-1} + 0.9\varepsilon_{t+\tau-2} + 0.8\varepsilon_{t+\tau-3}$ 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 $\tau = 1$, $\gamma = [0.15\ 0.15\ 0.75]'$ 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:

$$x_{i,t} = a_i + \delta_i x_{i,t-1} + \nu_{it} \tag{16}$$

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 time 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 described in (15), but the vector x_t follows the VAR(1) process:

$$x_t = a + \Phi x_{t-1} + v_t$$

with $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$, a a 3×1 vector of ones, $u_{t+\tau}$ a $MA(\tau - 1)$ process of the same form as for DGP1, $v_t \sim N(0, I)$ and $u_{t+\tau}$ independent of v_t. The VAR(1) regression coefficient Φ allows for interdependence between the predictors:

$$\Phi = \begin{bmatrix} 0.6 & 0.1 & 0 \\ 0.6 & 0.25 & 0 \\ 0 & 0 & 0.9 \end{bmatrix}. \tag{17}$$

[18] The MA process for the errors is taken from Clark and McCracken (2012).

In the size exercise the vector γ is set to zeros, while in the power exercise $\gamma = [0.15\ 0.15\ -0.15]'$ 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 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:

$$M_0 : y_{t+\tau} = c_0 + \beta_0 y_t + u_{0,t+\tau}. \tag{18}$$

while the alternative models take the form:

$$M_m : y_{t+\tau} = \beta_m' x_{m,t} + u_{m,t+\tau},\ m = 1, ..., M, \tag{19}$$

where $x_{m,t} = (1, y_t, x_{1,t})'$ for $m = 1$, $x_{m,t} = (1, y_t, x_{1,t}, x_{2,t})'$ for $m = 2$, $x_{m,t} = (1, y_t, x_{1,t}, x_{2,t}, x_{3,t})'$ 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 2.2, where the alternative models are progressively nested within each other.

For both DGPs the estimates are carried out through OLS with expanding window scheme and 10 percent significance level. For the QLR test the simulations are implemented with $\hat{W} = \hat{V}^{-1}$, where $\hat{V}$ is defined as in (11). We focus on one and four step ahead forecasts and we consider different length of the in-sample, $T = \{80, 100, 200\}$, 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.

5.2 Simulation Results

5.2.1 Size

First we report the empirical size for the QLR, the max-t and the max-F statistic based on both adjusted (Table 1) and unadjusted MSPEs (Table 2). We also include the superior predictive ability tests by White (2000), Hansen (2005) and Corradi and Swanson (2007) described in Section 3.3. We abbreviate them as SPA_W, SPA_H and SPA_CS respectively.

Table 1. Empirical Size (Nominal Size =10%)

		DGP1						DGP2					
$\tau = 1$													
		P=40			P=100			P=40			P=100		
test	T	80	100	200	80	100	200	80	100	200	80	100	200
Bootstrap													
QLR_I		0.095	0.083	0.096	0.103	0.098	0.096	0.113	0.109	0.116	0.100	0.101	0.104
QLR_D		0.086	0.078	0.091	0.085	0.085	0.082	0.101	0.102	0.125	0.095	0.109	0.110
max-t		0.088	0.075	0.090	0.089	0.077	0.089	0.132	0.127	0.138	0.114	0.120	0.130
max-F		0.090	0.086	0.085	0.088	0.082	0.081	0.145	0.124	0.140	0.125	0.115	0.132
Normal													
QLR_I		0.129	0.135	0.135	0.099	0.101	0.106	0.126	0.129	0.133	0.097	0.098	0.112
QLR_D		0.095	0.093	0.103	0.062	0.066	0.072	0.095	0.101	0.104	0.065	0.066	0.074
max-t		0.065	0.068	0.074	0.052	0.052	0.058	0.073	0.071	0.079	0.053	0.056	0.059
Non-nested													
SPA_W		0.017	0.024	0.024	0.006	0.005	0.018	0.022	0.020	0.035	0.007	0.005	0.017
SPA_H		0.026	0.021	0.029	0.005	0.007	0.020	0.022	0.025	0.035	0.008	0.004	0.017
SPA_CS		0.179	0.153	0.121	0.319	0.245	0.158	0.037	0.031	0.009	0.069	0.041	0.023
$\tau = 4$													
Bootstrap													
QLR_I		0.167	0.188	0.203	0.183	0.183	0.169	0.132	0.125	0.147	0.128	0.135	0.151
QLR_D		0.150	0.168	0.200	0.160	0.162	0.143	0.123	0.137	0.153	0.130	0.158	0.133
max-t		0.144	0.135	0.156	0.141	0.148	0.123	0.147	0.161	0.178	0.128	0.171	0.147
max-F		0.320	0.304	0.334	0.295	0.325	0.295	0.240	0.237	0.245	0.215	0.245	0.225
Normal													
QLR_I		0.336	0.344	0.338	0.218	0.217	0.230	0.304	0.292	0.284	0.187	0.185	0.199
QLR_D		0.272	0.284	0.273	0.163	0.157	0.179	0.258	0.246	0.242	0.147	0.142	0.159
max-t		0.164	0.166	0.161	0.112	0.109	0.121	0.141	0.132	0.133	0.088	0.086	0.101
Non Nested													
SPA_W		0.117	0.125	0.143	0.029	0.058	0.092	0.116	0.121	0.123	0.043	0.052	0.091
SPA_H		0.158	0.154	0.187	0.051	0.073	0.107	0.137	0.137	0.151	0.057	0.061	0.108
SPA_CS		0.214	0.206	0.168	0.348	0.291	0.196	0.092	0.078	0.038	0.181	0.129	0.062

NOTE: The DGPs are described in Section 5.1. 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 '-F' refers to a test statistic constructed using the variance-covariance matrix of the forecast errors. The QLR_I and QLR_D statistics accommodate different structures of the alternative models as described in Section 2. The reported results are based on 10000 replications when the statistics are obtained under the assumption of normality. For critical values generated through the bootstrap the number of replications is 1000 and for each replication we generate 500 bootstrap samples. For every draw, the initial 100 observations generated are discarded. The superior predictive ability (SPA) tests are described in Section 3.3.

Tests for comparisons where all alternative models nest the benchmark and that are based on adjusted MSPEs exhibit reasonable size properties. When bootstrap critical values are employed all tests are slightly undersized for DGP1 and show good size properties also for DGP2 for τ=1. The max-t test is slightly oversized in that case, and max-F is most

oversized for P=40. For P=100 the QLR_I test performs best in terms of size.[19]

For the normal critical values the size distortions of tests based on adjusted MSPEs for τ=1 are larger than for those based on bootstrap critical values, which is in line with results in CM. The QLR_I is oversized and QLR_D is well sized to undersized, while the max-t is somewhat undersized (the latter finding confirms results in HW and CM for different DGPs). QLR_D has the best size of the three tests with normal critical values for P=40 and QLR_I has the best size for P=100. For four steps ahead and bootstrapped critical values all tests are oversized, but the max-F test shows the largest size distortions with empirical size even three times larger than the nominal size. For DGP 1 the max-t test does best in terms of size, while for DGP 2 the QLR_I test is best. With critical values based on the normal approximation the max-t is somewhat oversized, but has good size properties for higher values of P. In contrast, QLR_I and QLR_D are clearly oversized in this case, especially QLR_I for P=40.[20] We stress that, because the limiting distribution of the MSPE differentials is non-normal, the approximation is not guaranteed to provide well sized tests in a broader set of DGPs.

The tests of superior predictive ability (SPA) are suited for comparisons where at least one of the alternative models does not nest the benchmark. They are also constructed for a different null than the tests of equal predictive accuracy that we have discussed so far.Then it is not surprising that those tests do not perform as well as the tests designed for nested model comparison. In particular, we find that the SPA_W and SPA_H tests are both severely undersized for τ=1. Such large size distortions are in line with earlier simulation results for the White and Hansen tests in HW and CM. For τ=1 the SPA_CS test is oversized for DGP1, severely oversized for T=80, while for DGP2 it is undersized. For the SPA_CS there are, to our knowledge, no other simulation results published in the literature. For τ=4 SPA_W and SPA_H are oversized for P=40 and undersized for P=100. For SPA_H we find that the size increases with increasing in-sample period T, similar to the pattern found in CM. SPA_CS exhibits severe size distortions for DGP1, while for DGP2 the size is overall good, either under or oversized depending on the sample size. We have investigated further the properties of the SPA test by changing the forecasting models to consider only non-nested

[19] Simulation experiments in CM show that the bootstrap approach to obtain the critical values works well even for a large number of models when the number of predictors is relatively small.

[20] Further research might also explore the possibility of improving the small sample size properties of the QLR tests using different weighting matrix than $\hat{V}^{-1}$. In the appendix for example we show that for critical values obtained by simulating the asymptotic distribution, size improves in small samples when choosing a diagonal weighting matrix.

models and, in the case of the SPA_CS test, to introduce misspecification. The results in a Not-for-publication Appendix show improved size of the SPA tests in the simulation setting that considers nonnested models.

Table 2. Empirical Size (Nominal Size =10%), unadjusted statistics

		DGP1						DGP2					
		$\tau = 1$											
		P=40			P=100			P=40			P=100		
test	T	80	100	200	80	100	200	80	100	200	80	100	200
Bootstrap													
QLR_I-unadj		0.093	0.084	0.098	0.095	0.098	0.101	0.118	0.109	0.125	0.107	0.099	0.109
QLR_D-unadj		0.100	0.079	0.098	0.088	0.097	0.092	0.123	0.099	0.130	0.094	0.111	0.109
max-t-unadj		0.099	0.082	0.096	0.091	0.086	0.097	0.138	0.130	0.150	0.124	0.123	0.148
max-F-unadj		0.094	0.096	0.101	0.096	0.095	0.083	0.132	0.138	0.0145	0.127	0.124	0.151
Normal													
QLR_I-unadj		0.084	0.097	0.108	0.055	0.060	0.070	0.083	0.089	0.103	0.053	0.055	0.072
QLR_D-unadj		0.042	0.047	0.063	0.009	0.015	0.025	0.037	0.046	0.063	0.011	0.014	0.026
max-t-unadj		0.019	0.025	0.038	0.005	0.007	0.015	0.025	0.028	0.045	0.007	0.010	0.018
		$\tau = 4$											
Bootstrap													
QLR_I-unadj		0.161	0.171	0.191	0.162	0.171	0.159	0.144	0.149	0.167	0.142	0.135	0.156
QLR_D-unadj		0.149	0.171	0.188	0.147	0.156	0.146	0.133	0.155	0.185	0.137	0.159	0.144
max-t-unadj		0.130	0.147	0.166	0.117	0.127	0.112	0.142	0.169	0.188	0.143	0.175	0.164
max-F-unadj		0.264	0.245	0.295	0.217	0.228	0.229	0.207	0.221	0.231	0.167	0.190	0.203
Normal													
QLR_I-unadj		0.260	0.270	0.283	0.123	0.127	0.157	0.225	0.227	0.239	0.112	0.113	0.138
QLR_D-unadj		0.171	0.186	0.203	0.048	0.052	0.083	0.155	0.156	0.177	0.044	0.046	0.076
max-t-unadj		0.076	0.080	0.099	0.020	0.025	0.046	0.064	0.064	0.081	0.019	0.020	0.042

NOTE: Refer to the note on Table 1. Also, the suffix '-unadj' refers to test statistics constructed based on the differences in MSPE without using the CW adjustment.

Tests for nested model comparisons based on the unadjusted MSPEs (Table 2) show generally more size distortions than tests based on adjusted MSPEs. The size results show a similar pattern as for the adjusted statistics. Assuming the normal approximation leads to a severely undersized max-t test and undersized QLR tests for τ=1. For τ=4 our results show an undersized max-t test and somewhat over- or undersized tests for small values of T and small values of T and P respectively. The size distortions for the normal approximation in case of the unadjusted test statistic are not surprising since approximate normality has been shown to lead to reasonable small sample size properties in CW and HW only for the statistic based on adjusted MSPEs (see also the discussion in Section 4.3 in this paper). Regarding the max-F test we find similar relative performance as for the adjusted statistics, namely comparable size to QLR and max-t tests for τ=1 and much worse size for τ=4.

5.2.2 Power

We now comment on the power of the tests given the results shown in Table 3 and Table 4. For all tests and forecast horizons the power increases with the size of the out-of-sample period for a given in-sample size. Improvements are also generally obtained when the in-sample size grows for fixed out of sample size, but the power gain is more substantial for an increase in P than an equal increase in T.

Table 3. Power, tests based on adjusted MSPEs

		DGP1						DGP2					
		$\tau = 1$											
		P=40			P=100			P=40			P=100		
test	T	80	100	200	80	100	200	80	100	200	80	100	200
Bootstrap													
QLR_I		0.592	0.557	0.617	0.943	0.945	0.956	0.670	0.675	0.690	0.941	0.951	0.958
QLR_D		0.669	0.682	0.705	0.962	0.972	0.980	0.770	0.770	0.790	0.969	0.974	0.977
max-t		0.765	0.754	0.783	0.967	0.979	0.991	0.821	0.822	0.849	0.971	0.975	0.980
max-F		0.916	0.915	0.960	0.991	0.997	0.999	0.927	0.943	0.978	0.987	0.993	0.999
Normal													
QLR_I		0.634	0.629	0.655	0.935	0.941	0.957	0.724	0.733	0.771	0.972	0.975	0.981
QLR_D		0.680	0.679	0.711	0.952	0.957	0.969	0.784	0.797	0.836	0.985	0.984	0.990
max-t		0.662	0.675	0.718	0.951	0.957	0.975	0.769	0.783	0.822	0.984	0.983	0.991
Non-nested													
SPA_W		0.625	0.625	0.789	0.832	0.849	0.923	0.727	0.767	0.861	0.909	0.935	0.956
SPA_H		0.237	0.262	0.332	0.445	0.514	0.542	0.375	0.390	0.398	0.613	0.626	0.666
SPA_CS		0.639	0.649	0.771	0.886	0.887	0.894	0.142	0.126	0.109	0.262	0.236	0.224
		$\tau = 4$											
Bootstrap													
QLR_I		0.563	0.581	0.611	0.930	0.940	0.956	0.329	0.337	0.378	0.608	0.618	0.663
QLR_D		0.635	0.657	0.688	0.960	0.957	0.971	0.426	0.423	0.453	0.705	0.730	0.772
max-t		0.691	0.735	0.758	0.966	0.974	0.982	0.504	0.506	0.547	0.717	0.747	0.800
max-F		0.945	0.950	0.978	0.995	0.998	1.000	0.694	0.748	0.848	0.860	0.893	0.957
Normal													
QLR_I		0.766	0.773	0.789	0.946	0.953	0.966	0.567	0.587	0.616	0.720	0.739	0.777
QLR_D		0.787	0.799	0.817	0.959	0.962	0.974	0.604	0.620	0.668	0.767	0.788	0.837
max-t		0.738	0.752	0.794	0.955	0.961	0.977	0.467	0.481	0.534	0.691	0.701	0.764
Non-Nested													
SPA_W		0.669	0.687	0.838	0.873	0.894	0.936	0.499	0.541	0.652	0.597	0.653	0.755
SPA_H		0.557	0.553	0.659	0.730	0.761	0.799	0.433	0.476	0.465	0.488	0.531	0.594
SPA_CS		0.676	0.655	0.756	0.870	0.883	0.893	0.207	0.196	0.151	0.319	0.256	0.185

NOTE: Refer to the note on Table 1.

In the case of a multivariate one-sided alternative hypothesis there is no uniformly more powerful test, so the ranking between tests is likely to vary across different simulation designs.

However given the structure of the alternative models the QLR_D test should outperform the QLR_I test as the latter does not account for the particular ordering of the MSPE differences. This conjecture is confirmed by our simulations.

The power based on adjusted test statistics is presented in Table 3. The QLR and max-t tests based on bootstrap critical values have comparable power for P=100 given the size properties, while for small P max-t has better power. QLR_D is more powerful than QLR_I for both τ=1 and τ=4. The max-F shows higher power than the max-t and QLR tests, in particular for DGP2 for τ=4. However, the size of this test is clearly distorted for four step ahead predictions which suggests only a limited usefulness of this test statistic in practice. Using normal critical values, the QLR tests are more powerful than the max-t test for DGP1, while for DGP2 the max-t test is more powerful.

Table 4. Power, tests based on unadjusted MSPEs

		DGP1						DGP2					
		$\tau = 1$											
		P=40			P=100			P=40			P=100		
test	T	80	100	200	80	100	200	80	100	200	80	100	200
Bootstrap													
QLR_I-unadj		0.217	0.221	0.275	0.465	0.501	0.546	0.400	0.384	0.400	0.603	0.639	0.679
QLR_D-unadj		0.391	0.394	0.405	0.775	0.781	0.740	0.602	0.591	0.584	0.869	0.875	0.860
max-t-unadj		0.524	0.496	0.526	0.846	0.844	0.827	0.688	0.678	0.665	0.899	0.896	0.888
max–F-unadj		0.751	0.742	0.817	0.932	0.940	0.970	0.822	0.835	0.894	0.962	0.972	0.968
Normal													
QLR_I-unadj		0.216	0.215	0.252	0.366	0.387	0.438	0.327	0.339	0.391	0.484	0.502	0.586
QLR_D-unadj		0.239	0.241	0.285	0.414	0.431	0.496	0.375	0.394	0.455	0.573	0.593	0.673
max-t-unadj		0.208	0.225	0.289	0.387	0.431	0.517	0.303	0.322	0.386	0.531	0.547	0.634
		$\tau = 4$											
Bootstrap													
QLR_I-unadj		0.307	0.310	0.311	0.554	0.545	0.560	0.289	0.280	0.300	0.355	0.369	0.426
QLR_D-unadj		0.395	0.425	0.402	0.751	0.733	0.726	0.391	0.395	0.404	0.595	0.598	0.623
max-t-unadj		0.505	0.541	0.516	0.824	0.813	0.825	0.434	0.443	0.449	0.618	0.623	0.648
max–F-unadj		0.768	0.807	0.866	0.931	0.930	0.998	0.587	0.632	0.712	0.725	0.749	0.832
Normal													
QLR_I-unadj		0.408	0.416	0.456	0.477	0.506	0.563	0.373	0.386	0.427	0.318	0.344	0.412
QLR_D-unadj		0.410	0.426	0.477	0.507	0.537	0.595	0.370	0.395	0.456	0.352	0.379	0.472
max-t-unadj		0.337	0.356	0.424	0.471	0.519	0.605	0.208	0.231	0.286	0.233	0.253	0.341

NOTE: Refer to the note on Table 1.

SPA_W and SPA_H exhibit low power, which was expected since the tests are severely undersized. SPA_CS has power comparable to the max-t and QLR tests, but the evaluation

of this power performance has to take into account that the test is clearly oversized, in particular for DGP2.

The power of the tests for nested model comparison based on unadjusted MSPE is displayed in Table 4. As noted in CM tests based on unadjusted statistics have lower power than tests based on adjusted statistics. With the bootstrap based critical values the max-t test exhibits higher power than the QLR test, although the difference to QLR_D is not very large. Also in this case the QLR_D test has larger power than QLR_I for τ=1 and τ=4. In contrast, with critical values obtained assuming normal approximation the QLR_D test has higher power than the max-t and QLR_I tests for both τ=1 and τ=4.

6 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 instability not 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. 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 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.[21] 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 5 collects test results for the max t-statistic and the two quasi likelihood ratio tests for both one step ahead and four step ahead forecasts. 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).

Table 5. Test of Equal Forecast Accuracy for US Inflation

	One − Step Ahead						*Four − Step Ahead*					
	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
max-t	1.744	0.655	0.106	1.145	0.557	0.247	2.052	0.056	0.086	1.109	0.230	0.260
QLR_I	4.300	0.545	0.121	2.487	0.573	0.340	4.349	0.094	0.146	2.020	0.363	0.439
QLR_D	4.301	0.536	0.078	0.712	0.589	0.538	4.350	0.082	0.123	0.143	0.612	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 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 only QLR_D rejects 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 are bootstrapped. For four step-ahead forecasts there is strong evidence against the null for the max-t stat and the QLR tests for bootstrapped critical values, while for critical values under normality only the max-t test rejects.[22]

We turn to our simulation results to interpret the findings in Table 5. For a four step ahead forecast horizon the max-t test for DGP2 (which is more relevant for our empirical

[21] Hubrich (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.

[22] Concerns may rise 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.

application) with both bootstrap and normal critical values exhibits higher power than the QLR tests. Also, the QLR_D test, which has higher power than QLR_I according to our simulations, also clearly rejects based on bootstrap critical values, and is close to rejection when using normal critical values. In light of these considerations we conclude that we can reject equal forecast accuracy for the 1st sample on a 10% nominal significance level for four-step ahead forecasts.

7 Conclusions

This paper introduces a quasi likelihood ratio 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 nonstandard and it depends on the characteristics of the predictors. Then we prove the validity of the bootstrap procedure developed in CM for our proposed test.

A further contribution of this paper is to discuss the tests of equal and superior predictive accuracy for multiple model comparisons suggested in the literature in a unified notational framework.

The finite sample size and power properties of the tests are evaluated and compared via Monte Carlo simulations either by bootstrapping or by treating the statistics as normally distributed. These investigations indicate that the bootstrapped critical values deliver QLR tests with empirical size close to nominal size for one-step-ahead forecasts, whereas for longer forecast horizons the tests are somewhat oversized, with the max-F suffering the largest distortions. The normal approximation of the vector of MSPE-adjusted yields approximately correctly sized QLR, max-t and max-F tests for one step ahead forecasts but for longer horizons it provides oversized QLR tests while the max-t test is about correctly sized. Relying on the bootstrap rather than on the assumption of normality to compute the critical values in general does not affect the power of the tests. Also, we find that the CW adjustment improves size in particular for higher forecast horizons, and power of the QLR tests, max-t and max-F tests at all forecast horizons. The size and power properties of the QLR tests relative to the max t-statistic depends on parameterization of the Monte Carlo experiment. We compare our test with existing tests for multi-model forecast comparison of superior predictive ability such as White (2000), Hansen (2005) and Corradi and Swanson (2007), which are suited for nonnested model comparison. These tests have inferior size properties

in our simulation setting which involves a different null and a nested model comparison.

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] Andrews, D. W. K. (2001), 'Testing when a Parameter is on the Boundary of the Maintained Hypothesis', Econometrica, vol. 69, 683-734.

[3] 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.

[4] Ashley, R., C.W.J. Granger and R.Schmalense (1980), 'Advertising and Aggregate Consumption: an Analysis of Causality', Econometrica, vol.48 n.5.

[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 (2005a), 'Evaluating Direct Multistep Forecasts', Econometric Reviews, vol.24, 369-404.

[9] Clark, T.E. and M. W. McCracken (2005b), 'The Power of Tests of Predictive Ability in the Presence of Structural Breaks', Journal of Econometrics, vol.124, 1-31.

[10] Clark, T.E. and M. W. McCracken (2012), 'Reality Checks and Comparisons of Nested Predictive Models', Journal of Business & Economic Statistics, vol. 30, 53-66.

[11] Clark, T.E. and M. W. McCracken (2011b), 'Advances in Forecast Evaluation', manuscript, St. Louis Federal Reserve Bank.

[12] 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.

[13] Clark, T.E. and K. West (2007), 'Approximately Normal Tests for Equal Predictive Accuracy in Nested Models', Journal of Econometrics, vol.138, 291-311.

[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, vol.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] Hansen, P. H. (2005), 'A Test for Superior Predictive Ability', Journal of Business and Economic Statistics, vol.23, 365-380.

[23] Harvey, D.I. and P. Newbold (2000), 'Tests for Multiple Forecast Encompassing', Journal of Applied Econometrics, vol.15, 471-482.

[24] Harvey, D.I., S.J. Leybourne and P. Newbold (2000), 'Tests for Forecast Encompassing', Journal of Business and Economic Statistics, vol.16, 254-259.

[25] 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.

[26] 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.

[27] Hubrich, K. and K. D. West (2010), 'Forecast Evaluation of Small Nested Model Sets', Journal of Applied Econometrics, vol.25, 574-594.

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

[29] McCracken, M. W. (2007), 'Asymptotics for Out of Sample Tests of Causality', Journal of Econometrics, vol. 140, 719-752.

[30] 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.

[31] Newey, W. K., West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', Econometrica, vol 55, 703-708.

[32] Perlman, M. D. (1969), 'One-Sided Testing Problems in Multivariate Analysis', The Annals of Mathematical Statistics, vol.40, 549-567.

[33] Ravazzolo, F. and P. Rothman (2010), 'Oil and US GDP: A Real-Time Out-of-Sample Examination', Journal of Money, Credit and Banking, forthcoming.

[34] Silvapulle M. J. and P. K. Sen (2005), 'Constrained Statistical Inference: Inequality, Order, and Shape Restrictions', Wiley-Interscience.

[35] Stock, J. H., and M. W. Watson (1999), 'Forecasting Inflation', Journal of Monetary Economics, vol. 22, 293-335.

[36] 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.

[37] Stock, J. H., and M. W. Watson (2010), 'Modeling Inflation After the Crisis', NBER wp 16488.

[38] West, K.D. (1996), 'Asymptotic Inference about Predictive Ability', Econometrica, vol.64, 1067-1084.

[39] West, K.D. (2006), 'Forecast Evaluation', in Handbook of Economic Forecasting, vol.1, 100-134, Elsevier.

[40] White, H. (1982), 'Maximum Likelihood Estimation of Misspecified Models', Econometrica, vol 50, 1-25.

[41] White, H. (1994), 'Estimation, Inference and Specification Analysis', Cambridge University Press.

[42] White, H. (2000), 'A Reality Check for Data Snooping', Econometrica, vol.68, 1097-1126.