Research Division
Federal Reserve Bank of St. Louis
Working Paper Series

Advances in Forecast Evaluation

Todd E. Clark
and
Michael W. McCracken

Working Paper 2011-025B
http://research.stlouisfed.org/wp/2011/2011-025.pdf

September 2011
Revised June 2012

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

Advances in Forecast Evaluation *

Todd E. Clark
Federal Reserve Bank of Cleveland

Michael W. McCracken
Federal Reserve Bank of St. Louis

June 2012

Abstract

This paper surveys recent developments in the evaluation of point forecasts. Taking West's (2006) survey as a starting point, we briefly cover the state of the literature as of the time of West's writing. We then focus on recent developments, including advancements in the evaluation of forecasts at the population level (based on true, unknown model coefficients), the evaluation of forecasts in the finite sample (based on estimated model coefficients), and the evaluation of conditional versus unconditional forecasts. We present original results in a few subject areas: the optimization of power in determining the split of a sample into in-sample and out-of-sample portions; whether the accuracy of inference in evaluation of multi-step forecasts can be improved with judicious choice of HAC estimator (it can); and the extension of West's (1996) theory results for population-level, unconditional forecast evaluation to the case of conditional forecast evaluation.

JEL Nos.: C53, C12, C52
Keywords: Prediction, equal accuracy

* *Clark*: Economic Research Dept.; Federal Reserve Bank of Cleveland; P.O. Box 6387; Cleveland, OH 44101; todd.clark@clev.frb.org. *McCracken* (corresponding author): Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. We gratefully acknowledged helpful comments from the editors, two reviewers, Kirstin Hubrich, and participants at a Federal Reserve Bank of St. Louis conference for the Handbook.

1 Introduction

Over time, many researchers have come to view forecast evaluation as a vital component of empirical time series work. Since at least the work of Fair and Shiller (1989, 1990) and Meese and Rogoff (1983, 1988), forecast evaluation has become an important metric for evaluating models. If one model is superior to another, it ought to forecast more accurately. Of course, forecast evaluation has long been important to applied forecasting. Forecasts need to be good to be useful for decision making. Determining if forecasts are good involves formal evaluation of the forecasts.

Since roughly the mid-1990s, the literature on forecast evaluation has mushroomed, in a variety of directions. In the first volume of the *Handbook of Economic Forecasting*, West (2006) provided a comprehensive survey of the extant literature. In this second volume, this chapter provides an update, focusing on developments in forecast evaluation since the time of West's writing. For that purpose, to put recent work in a broader context, we need to briefly cover some earlier developments, overlapping with some portions of West's survey. In this material, we extend West's overview for practitioners by including in an appendix a brief exposition of the derivations of some of the key results in the literature. We then focus on more recent developments, such as methods for evaluating population-level versus finite-sample forecast accuracy and the evaluation of conditional versus unconditional forecasts.

In this chapter, we also hone in on two outstanding issues in the literature, and present some original results on these issues. The first is obtaining accurate inference in evaluation of small samples of multi-step forecasts. The second issue is the optimization of power in determining the split of a sample into in-sample and out-of-sample portions. We provide a Monte Carlo assessment of options — alternative estimators of heteroskedasticity-and-autocorrelation (HAC) consistent variances — for obtaining small-sample inferences more reliable than those evident from some prior Monte Carlo work. We also present some original analysis extending West's (1996) results to include conditional forecasts.

We should note up front that, throughout the chapter, we focus on the evaluation of point forecasts. For overviews of the literature on the evaluation of density forecasts, we refer the reader to the comprehensive survey of Corradi and Swanson (2006) and the chapter by Andrew Patton in this volume, as well as the recent study of Chen (2011).

Our chapter proceeds as follows. Section 2 presents notation used throughout the chapter to represent the modeling and forecasting framework. Other, more specialized notation

is introduced as the chapter proceeds and the need for the notation arises. To reduce clutter, throughout the chapter our general approach is to define terms only once; to make notation easy to find, Table 1 provides a listing of notation used across multiple sections of the chapter. Section 3 reviews developments in the evaluation of pairs of forecasts, drawing a distinction between evaluation of population-level predictive ability and evaluation of finite-sample predictive ability. Section 4 reviews approaches to unconditional versus conditional forecast evaluation and includes our new extension of West's (1996) results from unconditional to conditional forecasts. Section 5 summarizes recent developments in methods for evaluating forecasts from multiple models. Section 6 reviews existing approaches to evaluating forecasts from models estimated with real-time data. To illustrate the use of some of the key tests of equal forecast accuracy, we end each of sections 3-6 with applications to forecasting inflation in the U.S.

Section 7 takes up small sample properties of the testing methods reviewed in previous chapters, first summarizing existing findings and then presenting our new Monte Carlo comparison of alternative HAC estimators in nested model forecast evaluation. Section 8 examines issues in the choice of the split of the sample into in-sample and out-of-sample portions, presenting our new results on power, and includes an overview of recent work on methods for testing across multiple sample splits. Section 9 discusses rationales for evaluating out-of-sample forecasts. Section 10 concludes with a brief summary. The Appendix provides some examples of the mathematics behind out-of-sample inference.

2 Modeling and Forecasting Framework

The sample of observations $\{y_t, x_t'\}_{t=1}^{T}$ includes a scalar random variable y_t to be predicted, as well as a $(k \times 1)$ vector of predictors x_t. Specifically, for each time t the variable to be predicted is $y_{t+\tau}$, where τ denotes the forecast horizon. The sample is divided into in-sample and out-of-sample portions. The total in-sample observations (on y_t and x_t) span 1 to R. Letting $P - \tau + 1$ denote the number of τ-step-ahead predictions, the total out-of-sample observations span $R + \tau$ through $R + P$. The total number of observations in the sample is $R + P = T$.[1]

[1] Our notation for the dependent variable y, and vector of predictors x, while standard in most econometrics textbooks, is not always sufficient for applied forecasting. For many macroeconomic variables (such as GDP) the forecasting agent actually has access to a triangular array of vintages of both the y's and x's. Put another way, in this and the next we chapters, we abstract from real-time data and the potential of having multiple vintages of data on a given variable. We take up real-time data in section 6.

The literature is largely silent on the best way to split the sample into in- and out-of-sample portions. There is, however, a clear trade-off. More out-of-sample observations (larger P) imply more forecasts and therefore more information regarding the accuracy of the forecasts. The converse is that more in-sample observations (larger R) imply that the parameter estimates will be more accurately estimated and likely lead to more accurate forecasts. As seen below, asymptotic inference on predictive ability often depends explicitly on the relative sample sizes, P/R. Section 8 considers in more detail the optimal choice of sample split and reviews recently developed approaches to testing across a wide range of samples.

Given the sample split, forecasts of $y_{t+\tau}$, $t = R, \ldots, T-\tau$, are generated using parametric models of the form $y_{t+\tau} = g(x_t, \beta^*) + u_{t+\tau}$ for a known function $g(\cdot, \cdot)$ and unknown finite-dimensioned parameter vector β^*. These parameters are estimated using one of three distinct observation windows. Under the recursive scheme, the parameter vector is updated at each forecast origin $t = R, \cdots, T-\tau$ using all available information. For example, if NLLS is used to estimate the above model, we have $\hat{\beta}_t = \arg\min_\beta \sum_{s=1}^{t-\tau}(y_{s+\tau} - g(x_s, \beta))^2$. Under the rolling scheme, the parameters are also updated at each forecast origin but always using the same number of observations R in the window, as, for example: $\hat{\beta}_t = \arg\min_\beta \sum_{s=t-\tau-R+1}^{t-\tau}(y_{s+\tau} - g(x_s, \beta))^2$. In our final scheme — the fixed scheme — the parameters are estimated only once at the initial forecast origin and hence $\hat{\beta}_t = \hat{\beta}_R = \arg\min_\beta \sum_{s=1}^{R-\tau}(y_{s+\tau} - g(x_s, \beta))^2$.

Regardless of the sample window used, the parameter estimates and the predictors are used to construct forecasts $\hat{g}_{t+\tau}(x_t, \hat{\beta}_t) = \hat{y}_{t+\tau}$ of the dependent variable at each forecast origin. These in turn can be used to construct forecast errors $\hat{u}_{t+\tau} = y_{t+\tau} - \hat{y}_{t+\tau}$. Typically the accuracy of the forecasts is evaluated based on a known function of this forecast error. Table 2 provides a list of several of the most common measures of "accuracy," using our loose interpretation of the term. The first three measures are intended to evaluate the accuracy of a single model, whereas the remaining ones are better thought of as evaluating the accuracy of a model relative to another model. West (2006) provides further detail on many of these measures, including references to original sources.

Note that regardless of the measures of accuracy (from Table 2) of interest, each can be written in a general form as $f(y_{t+\tau}, x_t, \hat{\beta}_t) = f_{t+\tau}(\hat{\beta}_t)$.[2] The goal of tests of predictive

[2] When two models are involved, redefine $\hat{\beta}_t$ as the vector formed by stacking the parameter estimates from each of the two models so that $\hat{\beta}_t = (\hat{\beta}'_{1,t}, \hat{\beta}'_{2,t})'$.

Table 1. Key Notation

Data-related

y_t = scalar variable to be predicted
x_t = vector of predictors
with nested models, $x_{2,t} = (x'_{1,t}, x'_{w,t})'$, vector with $k = (k_1 + k_w)$ elements
τ = forecast horizon
$T = R + P$, P = # of 1-step ahead forecasts, R = in-sample size, $\hat{\pi} = P/R$

Model and forecast-related

β_i = coefficient vector for model i with predictors $x_{i,t}$
$u_{i,t+\tau}$ = population forecast error from model $i = y_{t+\tau} - x'_{i,t}\beta_i^*$
$\hat{u}_{i,t+\tau}$ = estimated forecast error from model $i = y_{t+\tau} - x'_{i,t}\hat{\beta}_{i,t}$
with nested models, $u_{t+\tau} \equiv u_{2,t+\tau}$

Orthogonality conditions and loss functions

$h_{t+\tau} = h_{t+\tau}(\beta^*)$ = orthogonality conditions used to estimate model parameters
with more than one model, $h_{i,t+\tau}(\beta_i) = (y_{t+\tau} - x'_{i,t}\beta_i)x_{i,t}$
$f(y_{t+\tau}, x_t, \hat{\beta}_t) = f_{t+\tau}(\hat{\beta}_t)$ = forecast loss function
$\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$
$\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$
$\widehat{cw}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \left(\hat{u}^2_{2,t+\tau} - (x'_{2,t}\hat{\beta}_{2,t} - x'_{1,t}\hat{\beta}_{1,t})^2\right)$
$\text{MSE}_i = (P - \tau + 1)^{-1} \sum_{t=R}^{T-\tau} \hat{u}^2_{i,t+\tau}$

Moments and other terms in asymptotics

$\pi = \lim_{P,R\to\infty} P/R$, $\lambda = (1 + \pi)^{-1}$
$Eu^2_{2,t+\tau} = \sigma^2_2$; with nested models, $Eu^2_{t+\tau} = \sigma^2$
Ω = asymptotic variance of loss differential in West (1996)
$B = (Ex_t x'_t)^{-1}$; with nested models, $B_i = (Ex_{i,t}x'_{i,t})^{-1}$
$H(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$ (recursive scheme); with nested models, $H_2(t) = t^{-1}\sum_{j=1}^{t-\tau} h_{2,j+\tau}$
$F = E[\partial f_{t+\tau}(\beta)/\partial\beta]_{\beta=\beta^*}$
$S_{ff} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*))$
$S_{hh} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$
$S_{fh} = \lim_{T\to\infty} Cov(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*), T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$
$S_{\hat{f}\hat{f}} = \lim_{P\to\infty} Var((P - \tau + 1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - Ef_{t+\tau}(\hat{\beta}_t)))$
$J = (I_{k_1\times k_1}, 0_{k_1\times k_w})'$, $J_w = (0_{k_w\times k_1}, I_{k_w\times k_w})'$
$F_2 = J'_w B_2 J_w$
$\tilde{A}$ = a $(k_w \times k)$ matrix satisfying $\tilde{A}'\tilde{A} = B_2^{-1/2}(-J'B_1J + B_2)B_2^{-1/2}$
$\tilde{h}_{t+\tau} = \sigma^{-1}\tilde{A}B_2^{1/2}h_{2,t+\tau}$, $\tilde{H}_2(t) = \sigma^{-1}\tilde{A}B_2^{1/2}H_2(t)$
$\Gamma_{\tilde{h}\tilde{h}}(i) = E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-i}$
$S_{\tilde{h}\tilde{h}}$ = long-run variance of $\tilde{h}_{t+\tau} = \Gamma_{\tilde{h}\tilde{h}}(0) + \sum_{i=1}^{\tau-1}(\Gamma_{\tilde{h}\tilde{h}}(i) + \Gamma'_{\tilde{h}\tilde{h}}(i))$

Table 1, continued.

Test statistics

MSE-$t = \left((P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} \hat{d}_{t+\tau} \right) / \hat{S}_{dd}^{1/2}$, $\hat{S}_{dd}$ = long-run variance of $\hat{d}_{t+\tau}$

MSE-$F = \left(\sum_{t=R}^{T-\tau} \hat{d}_{t+\tau} \right) / \hat{\sigma}_2^2$, $\hat{\sigma}_2^2 = (P-\tau+1)^{-1} \sum_{t=R}^{T-\tau} \hat{u}_{2,t+\tau}^2$

ENC-$t = \left((P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} \hat{c}_{t+\tau} \right) / \hat{S}_{cc}^{1/2}$, $\hat{S}_{cc}$ = long-run variance of $\hat{c}_{t+\tau}$

ENC-$F = \left(\sum_{t=R}^{T-\tau} \hat{c}_{t+\tau} \right) / \hat{\sigma}_2^2$

Distributional terms

$W(\omega)$ = a $k_w \times 1$ vector standard Brownian motion

$\Gamma_1 = \int_\lambda^1 \omega^{-1} W'(\omega) S_{\tilde{h}\tilde{h}} dW(\omega)$

$\Gamma_2 = \int_\lambda^1 \omega^{-2} W'(\omega) S_{\tilde{h}\tilde{h}} W(\omega) d\omega$

$\Gamma_3 = \int_\lambda^1 \omega^{-2} W'(\omega) S_{\tilde{h}\tilde{h}}^2 W(\omega) d\omega$

$\Gamma_4 = \int_\lambda^1 (\vartheta' B_2^{-1/2} \tilde{A}'/\sigma) S_{\tilde{h}\tilde{h}}^{1/2} dW(\omega)$

$\Gamma_5 = (1-\lambda) \beta_w' F_2^{-1} \beta_w / \sigma^2$

Table 2. Common Measures of Point Forecast Accuracy

measure	$f_{t+\tau}(\beta)$
1. bias (zero mean prediction error)	$u_{t+\tau}$
2. serial correlation (zero first-order correlation)	$u_{t+\tau} u_{t+\tau-1}$
3. efficiency (no correlation between error and prediction)	$u_{t+\tau} g(x_t, \beta)$
4. encompassing (no correlation between model 1's error and model 2's prediction)	$u_{1,t+\tau} g_2(x_t, \beta)$
5. mean square error	$u_{t+\tau}^2$
6. mean absolute error	$\lvert u_{t+\tau} \rvert$
7. linex loss	$e^{\alpha u_{t+\tau}} - \alpha u_{t+\tau} - 1$

ability is to determine how best to use $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ as a means of telling us something about the unknown future accuracy of the model(s), as well as model adequacy.

3 Pairs of Models: Population-Level and Finite-Sample Inference

Starting with West (1996), much of the literature on forecast evaluation has focused on developing methods for testing population-level predictive ability, which involves using $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ to learn something about $Ef_{t+\tau}(\beta^*)$ — that is, the accuracy of the forecasts at unknown population values of parameters. Put another way, tests of population-level predictive ability are designed for evaluating the adequacy and accuracy of models if one had an infinite sample of data to estimate model parameters.

In a comparison of forecasts from nested models, tests of population-level predictive ability are effectively equivalent to tests of whether the additional parameters in the larger of the two models are zero. As a consequence, in a comparison of forecasts from nested models, a null of equal mean square error (MSE) can be rejected even though, in the finite sample at hand, the smaller model has a lower MSE than the larger model. This can occur because, in the finite sample, imprecision in parameter estimates can cause the MSE of the forecast from a true, larger model to exceed the MSE of the smaller model. The test rejection implies that, in a very large sample, the larger model would be estimated precisely enough that its forecasts could be expected to be more accurate than the forecasts from the smaller model.

In contrast, testing finite-sample predictive ability involves using $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ to learn something about $Ef_{t+\tau}(\hat{\beta}_t)$ — that is, the accuracy of the forecasts at estimated values of parameters. Put another way, tests of finite-sample predictive ability are designed to assess the accuracy of a model in a (finite) sample of the size at hand. In a comparison of forecasts from nested models, these tests can be seen as raising the bar relative to population-level tests: the question is not whether the additional coefficients of the larger model are zero (as in population-level tests), but are they non-zero and estimated accurately enough to make the competing models equally accurate in a finite sample? Under this approach, a null of equal MSE would only be rejected if, in the sample at hand, the rejected model's MSE exceeded the other model's MSE.[3]

[3] Our distinction between tests of population-level and finite-sample predictive ability directly parallels the concepts of "predictability" and "forecastability" in Hendry (2004).

This section first provides an overview of population-level forecast evaluation (relatively brief in light of the detail provided in West (2006)) and recent developments in population-level testing. Our presentation of population-level evaluation focuses on a limited set of tests of equal forecast accuracy, which have been the focus of the finite-sample evaluation literature and which have also been the source of new developments in population-level evaluation. West (2006) provides a comprehensive overview of a broader set of tests. Building on the population-level results, we then review three recently developed approaches to testing equal accuracy in the finite sample, due to Giacomini and White (2006), Clark and McCracken (2011a), and Calhoun (2011). The last subsection illustrates the use of some of the tests with an application to inflation forecasting. While section 3 focuses on providing an overview, the Appendix sketches the basics of the derivations of some key results in the literature.

3.1 Population-level predictive ability

For questions of population-level predictive ability, it is crucial that we recognize that $Ef_{t+\tau}(\beta^*)$ depends on β^*, the unknown true value of the parameter estimate $\hat{\beta}_t$. With this in mind, the original question can be recast as: Can $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ be used to learn something about the accuracy of the forecasts were we to know the true values of the model parameters?

3.1.1 Non-nested models

Building on earlier work by Diebold and Mariano (1995), West (1996) develops a theory for addressing this population-level question. In particular, he shows that

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-Ef_{t+\tau}(\beta^*)) \rightarrow^d N(0,\Omega), \tag{1}$$

and hence for a given null hypothesis regarding $Ef_{t+\tau}(\beta^*)$, asymptotically valid inference can be conducted using standard normal critical values so long as one can obtain an asymptotically valid estimate of Ω.[4]

The details of how to estimate Ω is perhaps the main technical development in West (1996). Before providing this result, some additional notation and assumptions are needed.[5]

[4] Studies such as Corradi and Swanson (2007) have developed bootstrap-based inference approaches that can be applied with tests that have power against generic alternatives or with tests applied to forecasts from misspecified models.

[5] These assumptions are intended to be expository, not complete. See West (1996) for more detail.

(A1) $\hat{\beta}_t = \beta^* + BH(t) + o_{a.s.}(1)$, where for some mean zero process $h_{t+\tau} = h_{t+\tau}(\beta^*)$ [with h denoting the orthogonality conditions used to estimate parameters, such as $h_{t+\tau} = x_t u_{t+\tau}$ for a single linear regression], $H(t)$ equals $t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$, $R^{-1}\sum_{s=t-R+1}^{t-\tau} h_{s+\tau}$, and $R^{-1}\sum_{s=1}^{R-\tau} h_{s+\tau}$ for the recursive, rolling, and fixed schemes, respectively, and B denotes a non-stochastic matrix.

(A2) The vector $(f_{t+\tau}(\beta^*), h'_{t+\tau})'$ is covariance stationary and satisfies mild mixing and moment conditions.[6]

(A3) $\lim_{P,R\to\infty} P/R = \pi$, a constant that is finite for the rolling and fixed schemes but can be infinite for the recursive scheme.

(A4) The vector $F = E[\partial f_{t+\tau}(\beta)/\partial\beta]_{\beta=\beta^*}$ is finite.[7]

(A5) Ω is positive definite.

Given these assumptions, West (1996) shows that the asymptotic variance Ω can take a variety of forms depending on how the parameters are estimated:

$$\Omega = S_{ff} + \lambda_{fh}(FBS'_{fh} + S_{fh}B'F') + \lambda_{hh}FBS_{hh}B'F', \tag{2}$$

where $S_{ff} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*))$, $S_{hh} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$, $S_{fh} = \lim_{T\to\infty} Cov(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*), T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$, and

	$\lambda_{fh} =$	$\lambda_{hh} =$
Recursive	$1 - \pi^{-1}\ln(1+\pi)$	$2(1 - \pi^{-1}\ln(1+\pi))$
Rolling, $\pi \leq 1$	$\pi/2$	$\pi - \pi^2/3$
Rolling, $1 < \pi < \infty$	$1 - (2\pi)^{-1}$	$1 - (3\pi)^{-1}$
Fixed	0	π

.

In equation (2) we see that Ω consists of three terms. The first, S_{ff}, is the long-run variance of the measure of accuracy when the parameters are known. The third term, $\lambda_{hh}FBS_{hh}B'F'$, captures the contribution of the variance due purely to the fact that we do not observe β^* but must estimate it instead. The second term, $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F')$, captures the covariance between the measure of accuracy and the estimation error associated with $\hat{\beta}_t$. Because the parameter estimates can be constructed using three different observation windows (recursive, rolling, and fixed) it is not surprising that the terms that arise due to estimation error depend on that choice via the terms λ_{fh} and λ_{hh}.

[6] Like most of the literature, West's (1996) asymptotics treat the forecast model size as fixed and finite. Anatolyev (2007) shows, using a fixed estimation scheme and West-type asymptotics, that allowing the size of the model to expand with the estimation and forecasting sample can greatly complicate the asymptotic distribution of tests of predictive ability.

[7] McCracken (2000) weakens this assumption to $F = \partial E[f_{t+\tau}(\beta)]/\partial\beta_{\beta=\beta^*}$ so that the function $f_{t+\tau}(\beta)$ need not be differentiable.

With this formula in hand, estimating Ω is straightforward. Since $\hat{\pi} = P/R \to \pi$ and both λ_{fh} and λ_{hh} are continuous in π, substituting $\hat{\pi}$ for π is sufficient for estimating both λ_{fh} and λ_{hh}. The F term can be estimated directly using $\hat{F} = (P - \tau + 1)^{-1} \sum_{t=R}^{T-\tau} \partial f_{t+\tau}(\hat{\beta}_t)/\partial\beta$.[8] When only one model has been estimated, the B term is typically the inverse of the Hessian matrix associated with the loss function used to estimate the model parameters. For example, if NLLS is used to estimate the model such that $\hat{\beta}_t = \arg\min_\beta \sum_{s=1}^{t-\tau}(y_{s+\tau} - g(x_s, \beta))^2$, then a consistent estimate of B is given by $\hat{B} = (T^{-1} \sum_{s=1}^{T-\tau} \partial^2(y_{s+\tau} - g(x_s, \hat{\beta}_T))^2/\partial\beta\partial\beta')^{-1}$. If more than one model is being used to construct $f_{t+\tau}(\hat{\beta}_t)$ (so that $\hat{\beta}_t = (\hat{\beta}'_{1,t}, \hat{\beta}'_{2,t})'$), then B is the block diagonal matrix $diag(B_1, B_2)$ and hence a consistent estimate is $\hat{B} = diag(\hat{B}_1, \hat{B}_2)$.

For the long-run variances and covariances needed to compute the test statistic, West (1996) shows that standard kernel-based estimators are consistent. To be more precise, define $\bar{f} = (P - \tau + 1)^{-1} \sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$, $\hat{\Gamma}_{ff}(j) = (P - \tau + 1)^{-1} \sum_{t=R+j}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - \bar{f})(f_{t+\tau-j}(\hat{\beta}_{t-j}) - \bar{f})'$, $\hat{\Gamma}_{hh}(j) = T^{-1} \sum_{t=j+1}^{T-\tau} h_{t+\tau}(\hat{\beta}_t) h'_{t+\tau-j}(\hat{\beta}_{t-j})$ and $\hat{\Gamma}_{fh}(j) = (P - \tau + 1)^{-1} \sum_{t=R+j}^{T-\tau} f_{t+\tau}(\hat{\beta}_t) h'_{t+\tau-j}(\hat{\beta}_{t-j})$, with $\hat{\Gamma}_{ff}(j) = \hat{\Gamma}_{ff}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}'_{hh}(-j)$, and $\hat{\Gamma}_{fh}(j) = \hat{\Gamma}'_{fh}(-j)$. The long-run variance estimates $\hat{S}_{ff}$, $\hat{S}_{hh}$, and $\hat{S}_{fh}$ are then constructed by weighting the relevant leads and lags of these covariances, as in HAC estimators such as that developed by Newey and West (1987).

Interestingly, for some cases estimating Ω is as simple as using the estimate $\hat{\Omega} = \hat{S}_{ff}$. This arises when the second and third terms in equation (2), those due to estimation error, cancel and hence we say the estimation error is asymptotically irrelevant.

Case 1. If $\pi = 0$, then both λ_{fh} and λ_{hh} are zero and hence $\Omega = S_{ff}$. This case arises naturally when the sample split is chosen so that the number of out-of-sample observations is small relative to the number of in-sample observations. Chong and Hendry (1986) first observed that parameter estimation error is irrelevant if P is small relative to R.

Case 2. If $F = 0$, then $\Omega = S_{ff}$. This case arises under certain very specific circumstances but arises most naturally when the measure of "accuracy" is explicitly used when estimating the model parameters. The canonical example is the use of a quadratic loss function (MSE) to evaluate the accuracy of forecasts from two non-nested models estimated by ordinary or non-linear least squares. In this situation, the F term equals zero and estimation error is asymptotically irrelevant.

[8] If $f_{t+\tau}(\beta)$ is non-differentiable see McCracken (2004) for an alternative estimator.

Case 3. There are instances where $-S_{fh}B'F' = FBS_{hh}B'F'$ and hence under the recursive scheme, estimation error is asymptotically irrelevant. In this case, it isn't so much than any particular term equals zero but that the sum of the components just happens to cancel to zero. One such example is a test for zero mean prediction error in models that contain an intercept, for which the Appendix sketches the asymptotic derivations. See West (1996, 2006) and West and McCracken (1998) for other examples.

3.1.2 Nested models

Although the results in West (1996) have many applications, the theory is not universal. In particular, one of the primary assumptions for the results in West (1996) to hold is that Ω must be positive. In nearly all the examples from Table 2, this is not an issue. However, problems arise in applications where one wishes to compare the accuracy of two models that are nested under the null of equal population-level forecast accuracy. Consider the case where two nested OLS-estimated linear models are being compared. If we define the $(k \times 1, k = k_1 + k_w)$ vector of predictors $x_t = x_{2,t} = (x'_{1,t}, x'_{w,t})'$, the models take the form $y_{t+\tau} = x'_{i,t}\beta^*_i + u_{i,t+\tau}$, for $i = 1, 2$, such that model 2 nests model 1 and hence $\beta^*_2 = (\beta^{*\prime}_1, \beta^{*\prime}_w)' = (\beta^{*\prime}_1, 0)'$ under the null. If we use quadratic loss to measure accuracy, we find that $f_{t+\tau}(\beta^*) = (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 - (y_{t+\tau} - x'_{2,t}\beta^*_2)^2 = (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 - (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 = 0$ for all t. Put in words, in population, under the null, the forecast errors from the competing errors are exactly the same at all points in time. Hence, it is clearly the case that S_{ff}, S_{fh}, and F all equal zero, making Ω also equal zero.

In this case, Clark and McCracken (2001, 2005a) and McCracken (2007) develop a different set of asymptotics that allow for an out-of-sample test of equal population-level unconditional predictive ability between two nested models. The key to their theory is to note that while $P^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - 0) \to^p 0$ when the models are nested, $\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - 0)$ need not have a degenerate asymptotic distribution. Building on this insight they show that, in the context of linear, OLS-estimated, direct-multistep forecasting models, a variety of statistics can be used to test for equal forecast accuracy and forecast encompassing despite the fact that the models are nested. Let $\hat{u}_{i,t+\tau} = y_{t+\tau} - x'_{i,t}\hat{\beta}_{i,t}$, $i = 1, 2$, $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$, $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$, and $\hat{\sigma}^2_2 = (P - \tau + 1)^{-1}\sum_{t=R}^{T-\tau}\hat{u}^2_{2,t+\tau}$. If we let $\hat{S}_{dd}$ and $\hat{S}_{cc}$ denote long-run variance estimates for, respectively, $\hat{d}_{t+\tau}$ and $\hat{c}_{t+\tau}$ (analogous to $\hat{S}_{ff}$ above) constructed with a HAC estimator such as Newey and West's (1987), these statistics take the form

$$\text{MSE-}t = \frac{(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}}{\hat{S}_{dd}^{1/2}}, \quad \text{MSE-}F = \frac{\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}}{\hat{\sigma}_2^2} \tag{3}$$

$$\text{ENC-}t = \frac{(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}}{\hat{S}_{cc}^{1/2}}, \quad \text{ENC-}F = \frac{\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}}{\hat{\sigma}_2^2}. \tag{4}$$

With nested models and a null hypothesis of equal predictive ability in population, these tests are naturally conducted with one-sided alternatives. Ashley, Granger, and Schmalensee (1980) first suggested that tests of equal accuracy of forecasts from nested models should be one-sided. In the case of tests for equal MSE, the reasoning is straight-forward. Under the null that $x_{w,t}$ has no predictive power for $y_{t+\tau}$, the population difference in MSEs will equal 0. Under the alternative that $x_{w,t}$ has predictive power, the population difference in MSEs will be positive ($\text{MSE}_2 < \text{MSE}_1$). As a result, the MSE-$t$ and MSE-F tests are one–sided to the right.

The more-involved logic for one-sided tests of forecast encompassing (which applies to both non-nested and nested model comparisons) was first laid out in Harvey, Leybourne, and Newbold (1998). Under the null that $x_{w,t}$ has no predictive power for $y_{t+\tau}$, the population covariance between $u_{1,t+\tau}$ and $(u_{1,t+\tau} - u_{2,t+\tau})$ will equal 0 (with nested models, the population forecast errors of the models will be exactly the same). Under the alternative that $x_{w,t}$ does have predictive power, the covariance will be positive. To see why, consider the forecast combination regression $y_{t+\tau} = (1-\alpha)g_{1,t+\tau} + \alpha g_{2,t+\tau} + \text{error}$, where g_1 and g_2 denote forecasts from the restricted and unrestricted models, respectively. Subtracting $g_{1,t+\tau}$ from both sides, and making the substitution $u_{1,t+\tau} - u_{2,t+\tau} = g_{2,t+\tau} - g_{1,t+\tau}$, yields the encompassing regression $u_{1,t+\tau} = \alpha(u_{1,t+\tau} - u_{2,t+\tau}) + \text{error}$. If $x_{w,t}$ does have predictive power, such that model 2 is true, the population combination coefficient α equals 1. As a result, the covariance between $u_{1,t+\tau}$ and $(u_{1,t+\tau} - u_{2,t+\tau})$ will be positive. Consequently, the ENC-t and ENC-F tests are one–sided to the right.

Turning to asymptotic distributions, for each test the distributions have representations as functions of stochastic integrals of quadratics in Brownian motion. To illustrate essential features, we present selected results, for the distributions of the MSE-t and MSE-F tests when the recursive sampling scheme is used, developed in Clark and McCracken (2005a). The Appendix sketches the basics of the necessary derivations. These asymptotic results

require the following additional notation. Let (assume) $\lim_{P,R\to\infty} P/R = \pi \in (0,\infty)$, and define $\lambda = (1+\pi)^{-1}$. Let $h_{i,t+\tau}(\beta_i) = (y_{t+\tau} - x'_{i,t}\beta_i)x_{i,t}$, $h_{i,t+\tau} = h_{i,t+\tau}(\beta_i^*)$, and $Eu^2_{2,t+\tau} = Eu^2_{t+\tau} = \sigma^2$. For $H_2(t) = t^{-1}\sum_{j=1}^{t-\tau} h_{2,j+\tau}$, $B_i = (Ex_{i,t}x'_{i,t})^{-1}$ $i = 1,2$, the selection matrix $J = (I_{k_1\times k_1}, 0_{k_1\times k_w})'$, and a $(k_w \times k)$ matrix $\tilde{A}$ satisfying $\tilde{A}'\tilde{A} = B_2^{-1/2}(-J'B_1J + B_2)B_2^{-1/2}$, let $\tilde{h}_{t+\tau} = \sigma^{-1}\tilde{A}B_2^{1/2}h_{2,t+\tau}$ and $\tilde{H}_2(t) = \sigma^{-1}\tilde{A}B_2^{1/2}H_2(t)$. If we define $\Gamma_{\tilde{h}\tilde{h}}(i) = E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-i}$, then $S_{\tilde{h}\tilde{h}} = \Gamma_{\tilde{h}\tilde{h}}(0) + \sum_{i=1}^{\tau-1}(\Gamma_{\tilde{h}\tilde{h}}(i) + \Gamma'_{\tilde{h}\tilde{h}}(i))$. Finally, let $W(\omega)$ denote a $k_w \times 1$ vector standard Brownian motion, and define the following functionals: $\Gamma_1 = \int_\lambda^1 \omega^{-1}W'(\omega)S_{\tilde{h}\tilde{h}}dW(\omega)$, $\Gamma_2 = \int_\lambda^1 \omega^{-2}W'(\omega)S_{\tilde{h}\tilde{h}}W(\omega)d\omega$, and $\Gamma_3 = \int_\lambda^1 \omega^{-2}W'(\omega)S^2_{\tilde{h}\tilde{h}}W(\omega)d\omega$.

Under the assumptions of Clark and McCracken (2005a), it follows that

$$\begin{aligned} \text{MSE-}F &\to^d 2\Gamma_1 - \Gamma_2 \\ \text{MSE-}t &\to^d (\Gamma_1 - 0.5\Gamma_2)/\Gamma_3^{0.5}. \end{aligned} \tag{5}$$

These limiting distributions are neither normal nor chi-square when the forecasts are nested under the null. Hansen and Timmermann (2011) offer the following intuitive characterization of the MSE-F distribution. The first term (Γ_1) arises from the recursive estimation, with forecast errors mapping to $dW(\omega)$ and parameter estimation errors mapping to $W(\omega)$; the former influences the latter in later forecasts. The second term (Γ_2) stems from the accuracy loss associated with estimating more parameters in the larger model.

As the above equations suggest, the distributions generally depend upon the unknown matrix $S_{\tilde{h}\tilde{h}}$ that in turn depends upon the second moments of the forecast errors $u_{t+\tau}$, the regressors $x_{2,t}$, and the orthogonality conditions $h_{2,t+\tau}$. Algebraically, this dependence arises because, in the presence of conditional heteroskedasticity or serial correlation in the forecast errors, an information matrix-type equality fails: the expected outer product of the predictors is no longer proportional to the long run variance of $h_{2,t+\tau}$ with constant of proportionality σ^2. Similarly, in the context of likelihood-ratio statistics, Vuong (1989, Theorem 3.3) shows that the limiting distribution of the likelihood ratio statistic has a representation as a mixture of independent $\chi^2_{(1)}$ variates (in contrast to our integrals of weighted quadratics of Brownian motion). This distribution is free of nuisance parameters when the information matrix equality holds but in general does depend upon such nuisance parameters.

The limiting distributions are free of nuisance parameters if $S_{\tilde{h}\tilde{h}} = I$. If this is the case — if, for example, $\tau = 1$ and the forecast errors are conditionally homoskedastic

— the MSE-F representation simplifies to McCracken's (2007). Clark and McCracken (2005a) note that there is one other case in which the distributions of t-tests of equal MSE and forecast encompassing simplify to the nuisance parameter-free versions of Clark and McCracken (2001) and McCracken (2007): when $k_w = 1$, the scalar $S_{\tilde{h}\tilde{h}}$ can be factored out of both the numerator and denominator and hence cancels. Also, in the perhaps unlikely scenario in which each of the eigenvalues of $S_{\tilde{h}\tilde{h}}$ are identical, one can show that the limiting distributions no longer depend upon the value of $S_{\tilde{h}\tilde{h}}$.

When the limiting distribution is free of nuisance parameters, as in the case of forecast errors that are serially uncorrelated and exhibit conditional homoskedasticity, asymptotic critical values can be obtained from tables provided in Clark and McCracken (2001), McCracken (2007), and (in more detail) on these authors' webpages. These critical values were obtained by Monte Carlo simulations of the asymptotic distributions. These limiting distributions depend on two known parameters: the sample split parameter λ and the number of exclusion restrictions, k_w. As discussed in McCracken (2007), given λ, as k_w rises, the distribution of the MSE-F test drifts further into the negative orthant. Since the parameter λ enters the asymptotic distributions nonlinearly, its effect on their distributions is somewhat ambiguous. But we can say with certainty that the asymptotic mean of the MSE-F statistic decreases with λ just as it does with k_w.

For the cases in which the asymptotic distributions depend on unknown nuisance parameters that capture the presence of serial correlation in the forecast errors or conditional heteroskedasticity, Clark and McCracken (2005a) develop two alternative approaches to obtaining critical values. One approach is to compute asymptotic critical values from Monte Carlo simulations of the asymptotic distribution, which is a function of the variance matrix $S_{\tilde{h}\tilde{h}}$ that can be consistently estimated from the data. In the case of conditionally homoskedastic, one–step ahead forecast errors, the resulting critical values would be exactly the same as those of Clark and McCracken (2001) and McCracken (2007).

The second approach from Clark and McCracken (2005a) is to bootstrap data from a restricted VAR bootstrap, based on the parametric method of Kilian (1999). Under this bootstrap, vector autoregressive equations for y_t and x_t — restricted to impose the null that x has no predictive power for y — are estimated by OLS using the full sample of observations, with the residuals stored for sampling. Note that the DGP equation for y takes exactly the same form as the restricted forecasting model for $\tau = 1$ (but estimated

with all available data). In Clark and McCracken (2005a), in the case of the x equation, the lag orders for y and x are determined according to the AIC, allowing different lag lengths on each variable.[9] Bootstrapped time series on y_t and x_t are generated by drawing with replacement from the sample residuals and using the autoregressive structures of the VAR equations to iteratively construct data. In each bootstrap replication, the bootstrapped data are used to recursively estimate the restricted and unrestricted forecasting models — all specified in direct, multi-step form — on which the sample results are based. The resulting forecasts are then used to calculate forecast test statistics. Critical values are simply computed as percentiles of the bootstrapped test statistics.

While the asymptotic validity of the restricted VAR bootstrap for population-level forecast evaluation has not been established, it has been shown to work well in practice (e.g., Clark and McCracken (2001, 2005a), Clark and West (2006, 2007)). The primary hurdle in proving the validity of the bootstrap is the dependence of multi-step forecasts on non-linear functions of the parameters of the 1-step ahead VAR model. That is, the VAR in conventional 1-step ahead form implies multi-step forecasts that depend on polynomials of coefficients of the VAR. These non-linearities make it extremely difficult to prove the validity of the bootstrap. As described in section 3.1.4, more recent research has identified an alternative bootstrap approach for which validity can be proven.

For the ENC-t test applied to nested forecasting models, Clark and West (2006, 2007) show that, under certain conditions, the distribution is either asymptotically normal or approximately normal in practice. Clark and West demonstrate that the test can be viewed as an adjusted test for equal MSE, where the adjustment involves subtracting out of the difference in MSE a term that captures (under the null hypothesis of equal accuracy in population) the extra sampling error in the large model. Clark and West present the loss differential of the test statistic as

$$\widehat{cw}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \left(\hat{u}^2_{2,t+\tau} - (x'_{2,t}\hat{\beta}_{2,t} - x'_{1,t}\hat{\beta}_{1,t})^2\right),$$

where the correction term is the square of the difference in forecasts from the competing models. The average of this term over time captures the effect of additional parameter

[9] For the system of y, x equations to be used in the bootstrap, Clark and McCracken (2005a) adjust the coefficients of the OLS–estimated models for the small–sample bias that can plague time series models. Specifically, they use the bootstrap method proposed by Kilian (1998) to adjust the coefficients of the OLS–estimated models and then use the bias–adjusted forms as the bootstrap DGP equations. However, with the Monte Carlo designs and empirical applications we have considered, these bias adjustments don't usually have much effect on the resulting critical values or p-values.

estimation error in the larger model relative to the smaller. Because the difference in forecasts equals -1 times the difference in forecast errors, a little algebra shows that the loss differential $\widehat{cw}_{t+\tau}$ is 2 times the loss differential $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$ of the ENC-t test. Consequently, the t-statistic proposed by Clark and West (2006, 2007) is exactly the same as the ENC-t statistic.

Clark and West (2006) show that, in the special case of a null forecasting model that takes a martingale difference form (such as a no-change forecast implied by a random walk null, in which case the null model does not have estimated parameters), and alternative model forecasts generated with a rolling sample of data, the asymptotic distribution of the ENC-t test is standard normal. In the more general case of a null model that includes estimated parameters, Clark and West (2006, 2007) show that, within some limits on P/R and k_w settings (not necessarily all settings), the right-tail critical values can be reasonably approximated by standard normal critical values.

3.1.3 Overlapping models

Another situation for which the results of West (1996) do not apply arises when the models being compared are overlapping. Overlapping models is a concept introduced in Vuong (1989) in the context of comparing the relative fit of two (possibly) misspecified likelihood functions. For our purposes the concept is easier to present if we simply think of comparing two OLS-estimated linear regressions. Specifically, suppose we have two linear regressions that are intended to forecast excess returns r of some stock index

$$\begin{aligned} r_{t+1} &= \beta_{0,dy} + \beta_{dy} dy_t + u_{dy,t+1} \\ r_{t+1} &= \beta_{0,ep} + \beta_{ep} ep_t + u_{ep,t+1}, \end{aligned}$$

where dy and ep denote the corresponding dividend yield and earnings-price ratio, respectively. As Vuong notes, these two models can have equal predictive content two distinct ways. In the first, both β_{dy} and β_{ep} are non-zero and it happens to be the case that $E(u^2_{dy,t+1} - u^2_{ep,t+1}) = 0$. If this is the case we say the models are non-nested. In the second, both β_{dy} and β_{ep} are zero and hence $E(u^2_{dy,t+1} - u^2_{ep,t+1}) = 0$ but in the trivial sense that not only are the two models equally accurate but they are identical in population and hence $u_{dy,t+1} = u_{ep,t+1} \equiv u_{t+1}$. If this is the case we say the models are overlapping.

As Vuong notes, testing the null hypothesis that the two models are equally accurate (i.e., $E(u^2_{dy,t+1} - u^2_{ep,t+1}) = 0$) becomes much harder when one allows for the possibility that

the two models are overlapping. The problem is that the null hypothesis does not uniquely characterize the null asymptotic distribution. If the two models are non-nested the theory of West (1996), presented in section 3.1.1 for non-nested comparisons, can be used to show that for $\hat{d}_{t+1} = \hat{u}^2_{dy,t+1} - \hat{u}^2_{ep,t+1}$,

$$\text{MSE-}t = \frac{P^{-1/2}\sum_{t=R}^{T-1}\hat{d}_{t+1}}{\hat{S}_{dd}^{1/2}} \rightarrow^d N(0,1) \tag{6}$$

under the null hypothesis for some $\hat{S}_{dd} \rightarrow^p S_{dd} > 0$.

However, if the two models are overlapping we know $u_{dy,t+1} = u_{ep,t+1}$ and hence it must be that $S_{dd} = 0$. In this case the results of section 3.1.1 do not apply. In fact, Clark and McCracken (2012) show that the MSE-t statistic instead typically has a non-standard distribution akin to that derived for the case where two nested models are being compared. In the following we provide a brief description of these results.

Distribution of MSE-t for overlapping models Consider the case where two OLS-estimated linear models are being compared. The sample of observations $\{y_t, x_t'\}_{t=1}^T$ includes a scalar random variable y_t to be predicted, as well as a $(k_0+k_1+k_2 = k\times 1)$ vector of predictors $x_t = (x_{0,t}', x_{12,t}', x_{22,t}')'$. The two models are linear regressions with predictors $x_{1,t}$ and $x_{2,t}$ that share a common component $x_{0,t}$: $x_{1,t} = (x_{0,t}', x_{12,t}')'$ and $x_{2,t} = (x_{0,t}', x_{22,t}')'$.

Forecasts of y_{t+1}, $t = R, \ldots, T-1$, are generated using the two linear models $y_{t+1} = x_{1,t}'\beta_1^* + u_{1,t+1}$ and $y_{t+1} = x_{2,t}'\beta_2^* + u_{2,t+1}$. Under the null hypothesis of equal forecast accuracy between (degenerate) overlapping models, model 2 and model 1 collapse on one another for all t, and hence models $i = 1, 2$ include k_i excess parameters, respectively. Since this implies $\beta_i^* = (\beta_0^{*\prime}, 0')'$, the population forecast errors are identical under the null and hence $u_{1,t+1} = u_{2,t+1} \equiv u_{t+1}$ for all t. Both model 1's and model 2's forecasts are generated recursively using estimated parameters and hence models 1 and 2 yield two sequences of P forecast errors, denoted $\hat{u}_{1,t+1} = y_{t+1} - x_{1,t}'\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+1} = y_{t+1} - x_{2,t}'\hat{\beta}_{2,t}$, respectively.

Finally, the asymptotic results for overlapping models presented below use the following additional notation. Let $h_{t+1} = u_{t+1}x_t$, $H(t) = t^{-1}\sum_{s=1}^{t-1} h_{s+1}$, $B_i = (Ex_{i,t}x_{i,t}')^{-1}$, $B = (Ex_tx_t')^{-1}$, and $Eu_{t+1}^2 = \sigma^2$. For selection matrices

$$J_1' = \begin{pmatrix} I_{k_0\times k_0} & 0_{k_0\times k_1} \\ 0_{k_1\times k_0} & I_{k_1\times k_1} \\ 0_{k_2\times k_0} & 0_{k_2\times k_1} \end{pmatrix} \text{ and } J_2' = \begin{pmatrix} I_{k_0\times k_0} & 0_{k_0\times k_2} \\ 0_{k_1\times k_0} & 0_{k_1\times k_2} \\ 0_{k_2\times k_0} & I_{k_2\times k_2} \end{pmatrix} \tag{7}$$

and a $(k_1 + k_2 \times k)$ matrix $\tilde{A}$ satisfying $\tilde{A}'\tilde{A} = B^{-1/2}(-J_1'B_1J_1 + J_2'B_2J_2)B^{-1/2}$, let $\tilde{h}_{t+1} = \sigma^{-1}\tilde{A}B^{1/2}h_{t+1}$, $\tilde{H}(t) = \sigma^{-1}\tilde{A}B^{1/2}H(t)$ and $S_{\tilde{h}\tilde{h}} = E\tilde{h}_{t+1}\tilde{h}_{t+1}'$. Finally, let $W(\omega)$ denote a $(k_1 + k_2) \times 1$ vector standard Brownian motion, and define the following functionals: $\Gamma_1 = \int_\lambda^1 \omega^{-1}W'(\omega)S_{\tilde{h}\tilde{h}}dW(\omega)$, $\Gamma_2 = \int_\lambda^1 \omega^{-2}W'(\omega)S_{\tilde{h}\tilde{h}}W(\omega)d\omega$, and $\Gamma_3 = \int_\lambda^1 \omega^{-2}W'(\omega)S_{\tilde{h}\tilde{h}}^2W(\omega)d\omega$.

Under the assumptions of Clark and McCracken (2012), it follows that

$$\text{MSE-}t \rightarrow^d (\Gamma_1 - 0.5\Gamma_2)/\Gamma_3^{0.5}. \tag{8}$$

At first blush one might compare equation (8) with (5) and conclude that the distribution of the MSE-t statistic is the same whether we are comparing nested or overlapping models. While notationally they are identical, the distributions differ because the definitions of $\tilde{h}$ differ. Regardless, the distribution is non-standard and inference is made difficult due to the presence of the unknown matrix $S_{\tilde{h}\tilde{h}}$. Even so, Monte Carlo evidence suggests that a simple-to-implement fixed regressor wild bootstrap (discussed in section 3.1.4) can be used to construct asymptotically valid critical values when one knows the models are overlapping.

Testing Procedures Unfortunately, the theoretical results in the previous section are not particularly useful. They aren't useful because in practice one doesn't know whether the models are non-nested or whether the models are overlapping. Hence one doesn't know whether to use critical values associated with a standard normal distribution or whether to generate bootstrap-based critical values associated with the random variable in (8). In the following we delineate three possible testing provedures associated with the MSE-t statistic.

1. In the context of likelihood-ratio statistics, Vuong suggests a two-step procedure for testing the null hypothesis. In the first stage, a variance test, conducted at the α_1-percent level, is used to test the null that the population forecast errors are identical and hence the two models are overlapping. If we fail to reject, the procedure stops. Otherwise, if we reject the null (concluding that the two models are not overlapping), we conduct a test of equal accuracy at the α_2-percent level assuming the two models are non-nested. Vuong (1989) argues that this procedure controls the size of the test at the maximum of the nominal sizes used in each stage — i.e., controls max(α_1, α_2) — and hence the testing procedure is conservative.

This same logic extends to the use of out-of-sample statistics. As a corollary to the result in equation (8), Clark and McCracken (2012) show that the "variance" component of the MSE-t satisfies $P\hat{S}_{dd} \rightarrow^d 4\sigma^4\Gamma_3$ when the models are overlapping Moreover, their boot-

strap provides a method of estimating valid critical values associated with the asymptotic distribution of $P\hat{S}_{dd}$. As such our first testing procedure consists of (i) using the bootstrap to construct valid critical values associated with the distribution of $P\hat{S}_{dd}$. If we fail to reject the procedure stops. If we reject we then (ii) compare the $MSE-t$ statistic to standard normal critical values. If we fail to reject the procedure stops. Otherwise we reject the null hypothesis and conclude that the two models are not equally accurate. As was the case in Vuong (1989) this procedure is conservative and controls the nominal size of the test at the maximum of the nominal sizes (α_1 and α_2) used at each stage of the two-step procedure.

2. Alternatively one can construct a conservative one-step procedure. To see how this might be done note that the MSE-t statistic is bounded in probability regardless of whether the models are non-nested or overlapping. Define $q_{\alpha/2}$ and $q_{1-\alpha/2}$ as the lower and upper $\alpha/2$ percentiles of the MSE-t statistic when the models are overlapping. Define $z_{\alpha/2}$ and $z_{1-\alpha/2}$ as the same but when the models are non-nested – and hence these values correspond to percentiles associated with the standard normal distribution. With these percentiles in hand, a conservative test of the null hypothesis can be constructed by rejecting when the MSE-t statistic is less than $\min(q_{\alpha/2}, z_{\alpha/2})$ or when it is greater than $\max(q_{1-\alpha/2}, z_{1-\alpha/2})$. To use this procedure one needs access to the percentiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ but these can be estimated using the bootstrap discussed in the next section.

3. In a few isolated special cases, not previously discussed above, the MSE-t statistic is asymptotically standard normal even when the models are overlapping. As we found in our previous work on nested model comparisons, the MSE-t statistic is asymptotically standard normal when: (i) the number of out-of-sample forecasts P is small relative to the number of in-sample observations R used to estimate model parameters, such that $P/R \to 0$; or (ii) the fixed scheme is used to estimate model parameters and hence the parameters used for forecasting are not updated as we proceed across each forecast origin. When one of these two special cases is applicable, a two-step procedure is no longer necessary. We can test for equal forecast accuracy between two possibly overlapping models in just one step using standard normal critical values and still obtain an accurately sized test of equal accuracy.

3.1.4 Recent developments in population-level evaluation

Since West's (2006) survey, there have been two important extensions of the literature on evaluating pairs of forecasts at the population level, both for nested models. First, Hansen and Timmermann (2011) have extended the results of Clark and McCracken (2005a) and

McCracken (2007) by deriving a simplification of the asymptotic distribution of the MSE-F test, under less stringent assumptions. While Clark and McCracken (2005a) and McCracken (2007) use assumptions adapted from Hansen (1992), Hansen and Timmermann use assumptions based on de Jong and Davidson (2000), which are the weakest assumptions that can be used to ensure convergence to stochastic integrals. More importantly, Hansen and Timmermann are able to show that the asymptotic distribution of the MSE-F statistic simplifies to an eigenvalue-weighted average of a function (one for each eigenvalue of the matrix $S_{\tilde{h}\tilde{h}}$) of two independent χ^2-distributed random variables. In turn, with a 1-step ahead forecast horizon and conditional homoskedasticity of the forecast errors, the distribution sometimes simplifies to an analytical form. These simplifications offer the advantage of making asymptotic critical values easier to obtain, by eliminating the need for simulations in some cases, and make simulating critical values easier and more precise in general.

The second important extension is Clark and McCracken's (2011b) development of a fixed regressor bootstrap, which they prove to be asymptotically valid (and consistent) under assumptions similar to those of Clark and McCracken (2005a). Some researchers and practitioners may find it a little easier to implement than the restricted VAR bootstrap described above. The fixed regressor bootstrap's steps consist of the following.

1. (a) Use OLS to estimate the parameter vector β_1^* associated with the restricted model. Store the fitted values $x'_{1,s}\hat{\beta}_{1,T}$, $s = 1, \ldots, T-\tau$. (b) Use OLS to estimate the parameter vector β_2^* associated with the unrestricted model. Store the residuals $\hat{v}_{2,s+\tau}$, $s = 1, \ldots, T-\tau$.

2. If $\tau > 1$, use NLLS to estimate an $MA(\tau-1)$ model for the OLS residuals $\widehat{v}_{2,s+\tau}$ such that $v_{2,s+\tau} = \varepsilon_{2,s+\tau} + \theta_1\varepsilon_{2,s+\tau-1} + \cdots + \theta_{\tau-1}\varepsilon_{2,s+1}$.

3. Let η_s, $s = 1, \ldots, T$, denote an $i.i.d$ $N(0,1)$ sequence of simulated random variables. If $\tau = 1$, form a time series of innovations $\widehat{v}^*_{2,s+1} = \eta_{s+1}\widehat{v}_{2,s+1}$. If $\tau > 1$, form a time series of innovations computed as $\widehat{v}^*_{2,s+\tau} = (\eta_{s+\tau}\widehat{\varepsilon}_{2,s+\tau} + \widehat{\theta}_1\eta_{s+\tau-1}\widehat{\varepsilon}_{2,s+\tau-1} + \cdots + \widehat{\theta}_{\tau-1}\eta_{s+1}\widehat{\varepsilon}_{2,s+1})$, $s = 1, \ldots, T-\tau$.

4. Form artificial samples of $y^*_{s+\tau}$ using the fixed regressor structure, $y^*_{s+\tau} = x'_{1,s}\hat{\beta}_{1,T} + \widehat{v}^*_{2,s+\tau}$.

5. Using the artificial data, construct forecasts and an estimate of the test statistics (e.g., MSE-F, MSE-t, ENC-F, ENC-t) as if these were the original data.

6 Repeat steps 3-5 a large number of times: $j = 1, \ldots, N$.

7. Reject the null hypothesis, at the $\alpha\%$ level, if the test statistic is greater than the $(100-\alpha)\%$-ile of the empirical distribution of the simulated test statistics.

Finally, there has also been some significant progress in testing the rationality of a given forecast. West (2006) provides a summary of some of the key, previously existing methods for testing rationality. More recently, Patton and Timmermann (2012) develop new methods for testing the rationality or optimality of forecasts spanning multiple forecast horizons, using information from multiple horizons. Their methods exploit particular monotonicity properties of optimal forecasts. One such property is that, as the forecast horizon increases, the variance of an optimal forecast will decline. A second property is that optimal updating of forecasts implies that the variance of forecast revisions should be at least twice as large as the covariance between the revision and the actual value of the variable. Exploiting such monotonicity properties, Patton and Timmermann (2012) develop tests based on inequality constraints in a regression framework. They also develop versions of the optimality or rationality tests that use just forecasts, without need for data on the actual values of the variable being forecast. Monte Carlo evidence suggests the proposed tests improve on the size and power of conventional rationality tests. However, as some of the commentary published with the Patton and Timmermann (2012) article suggests (see, e.g., Hoogerheide, Ravazzolo, and van Dijk (2012) and West (2012)), there are some limitations to the proposed tests, with respect to application to forecasts from estimated models and forecasts from misspecified models. Still, the results of Patton and Timmermann (2012) represent an important step forward in methods for evaluating forecast rationality.

3.2 Finite-sample predictive ability

A test of finite-sample predictive ability addresses a different, but related, question than the one described in the previous subsection: Can we use $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ to learn something about $Ef_{t+\tau}(\hat{\beta}_t)$? For this question, it is crucial to recognize that $Ef_{t+\tau}(\hat{\beta}_t)$ depends on $\hat{\beta}_t$ and not the unknown true value of the parameter β^*. In other words, we want to know whether $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ can be used to learn something about the accuracy of the forecasts given that our forecasts are constructed using estimated parameters.

The importance of such a distinction is perhaps easiest to see when comparing the forecast accuracy of two nested models. Continuing with the notation above, we know that if $\beta_w^* = 0$, then the two models are identical and hence have equal population-level

predictive ability. We also know that if $\beta_w^* \neq 0$, then in population, the larger model will forecast more accurately than the smaller model. In practice, though, even when $\beta_w^* \neq 0$, the parameters are estimated with finite samples of data. It is then perfectly reasonable to consider the option that the smaller model is as accurate as (or even more accurate than) the larger model despite the fact that $\beta_w^* \neq 0$. This is particularly likely when the dimension of β_w^* is large relative to the existing sample size.

3.2.1 Giacomini and White (2006)

The first study to address this type of null hypothesis is Giacomini and White (2006). They note that two models can have equal forecast accuracy in finite samples if, continuing with our nested model comparison, the bias associated with estimating the misspecified restricted model happens to balance with the additional estimation error associated with estimating β_w^* in the correctly specified unrestricted model. This observation is perfectly true, but implementing a test for it is much harder, especially given a universe where you don't want to make extremely restrictive assumptions on the data (such as joint normality, conditionally homoskedastic and serially uncorrelated forecast errors, etc.). This scenario is much harder because we know in advance that any asymptotic approach to inference that allows the parameter estimates to be consistent for their population counterparts will imply that the unrestricted model is more accurate than the restricted model. In the notation of the tests of equal population-level predictive ability between finite-dimensioned nested models, this implies that any asymptotics that allow R to diverge to infinity will fail to be relevant for the null of equal finite-sample predictive ability.

As a result, Giacomini and White (2006) dispense with that assumption. More precisely they show that if the parameter estimates are constructed using a *rolling scheme with a finite observation window* R, then

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-Ef_{t+\tau}(\hat{\beta}_t)) \rightarrow^d N(0, S_{\hat{f}\hat{f}}), \tag{9}$$

where $S_{\hat{f}\hat{f}} = \lim_{P\rightarrow\infty} Var((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-Ef_{t+\tau}(\hat{\beta}_t)))$. Note that this differs from the asymptotic variance in West (1996) even when the second and third terms in Ω are asymptotically irrelevant since $S_{\hat{f}\hat{f}} \neq S_{ff}$.

This result is extremely powerful and covers a wide range of applications, including every example in Table 2. Interestingly, by requiring that the forecasts be constructed using a

small, finite, rolling window of observations, Giacomini and White (2006) are able to substantially weaken many of the assumptions needed for the results in Clark and McCracken (2001, 2005a), McCracken (2007), and West (1996). In particular, covariance stationarity of the observables is no longer needed — only that the observables are $I(0)$ with relatively mild mixing and moment conditions. There is no need for Ω to be positive (though $S_{\hat{f}\hat{f}}$ must be), and hence both nested and non-nested comparisons are allowed. The forecasts can be based on estimators that are Bayesian, nonparametric, or semi-parametric. The key is that R must be small and finite in all cases.

The primary weakness of the results in Giacomini and White (2006) is that their approach cannot be used with the recursive scheme. The recursive scheme fails because, absent any other assumptions on the finite dimensioned parameter β_w^*, as the sample size increases the parameter estimates $\hat{\beta}_t$ are consistent for their population counterparts and thus estimation error vanishes. Although the rolling scheme is relatively common among forecasting agents, it is by no means universal. Moreover, the asymptotics apply only when we think of the rolling observation window as small relative to the number of out-of-sample observations. Monte Carlo evidence on the magnitudes of P and R needed for accurate inference is limited. Most extant Monte Carlo work has focused on how small P/R needs to be make parameter estimation error asymptotically irrelevant, as opposed to how large the ratio needs to be for Giacomini and White asymptotics to be accurate.[10]

3.2.2 Clark and McCracken (2011a)

More recent work by Clark and McCracken (2011a) shows that, in some circumstances, one can construct a test of equal finite-sample unconditional predictive ability that permits not only the rolling scheme, but also the recursive scheme. In particular, they consider the case of testing this null hypothesis when comparing two nested OLS-estimated linear models and hence $Ef_{t+\tau}(\hat{\beta}_t) = E[(y_{t+\tau} - x_{1,t}'\hat{\beta}_{1,t})^2 - (y_{t+\tau} - x_{2,t}'\hat{\beta}_{2,t})^2] = 0$. The asymptotics are not unlike those from their previous work on equal population-level predictive ability (described in the previous section) but capture the bias and estimation error associated with, respectively, a misspecified restricted model and a correctly specified, but imprecisely estimated, unrestricted model.

But as noted above, since their results are asymptotic and the estimation error as-

[10] Clark and McCracken (2011c) consider larger P/R ratios than do most previous Monte Carlo assessments.

sociated with the parameter estimates vanishes asymptotically, balancing that estimation error with a bias component is problematic using standard parameterizations of a linear regression model. Instead Clark and McCracken (2011a) consider the case in which the additional predictors in the unrestricted model are "weak," using the following local-to-zero parameterization of the data generating process:

$$y_{t+\tau} = x'_{2,t}\beta^*_{2,T} + u_{t+\tau} = x'_{1,t}\beta^*_1 + x'_{w,t}(R^{-1/2}\beta^*_w) + u_{t+\tau}. \tag{10}$$

The intuition for this parameterization is based on an observation: As the sample size used to estimate the regression parameters increases, the estimation error associated with OLS estimation vanishes at a $\sqrt{T}$ rate. If bias due to model misspecification in the smaller (restricted) model is going to balance with the estimation error, it must also vanish at a $\sqrt{T}$ rate. To be clear, we do not take the model in equation (10) as a literal representation of the data, but rather consider it a tool for modeling how a bias-variance trade-off can exist in large samples as the size of the sample used for estimation increases.

As is the case for tests of equal population-level forecast accuracy between two nested models, the asymptotic distributions derived by Clark and McCracken (2011a) under weak predictability are nonstandard and have representations as functions of stochastic integrals of quadratics in Brownian motion. Moreover, the asymptotic distributions depend on unknown nuisance parameters that capture the presence of serial correlation in the forecast errors and conditional heteroskedasticity. Under the weak predictability null hypothesis, the nuisance parameters in the asymptotic distribution (under the null) also include the vector of coefficients on the weak predictors.

Consider, for example, the asymptotic distribution of the MSE-F test in equation (3). Under the assumptions of Clark and McCracken (2011a), the asymptotic distribution will depend on the stochastic integrals introduced in section 3.1.2 and the following: $\Gamma_4 = \int_\lambda^1 (\vartheta' B_2^{-1/2} \tilde{A}'/\sigma) S_{\tilde{h}\tilde{h}}^{1/2} dW(\omega)$ and $\Gamma_5 = (1-\lambda)\beta'_w F_2^{-1}\beta_w/\sigma^2$, where $J_w = (0_{k_w \times k_1}, I_{k_w \times k_w})'$, $\vartheta = (0_{k_1 \times 1}, \beta_w)'$, and $F_2 = J'_w B_2 J_w$. The asymptotic distribution is:

$$\text{MSE-}F \rightarrow^d \{2\Gamma_1 - \Gamma_2\} + 2\{\Gamma_4\} + \{\Gamma_5\}. \tag{11}$$

The first two terms of the asymptotic distribution (involving Γ_1 and Γ_2) are the same as in equation (5), which is the Clark and McCracken (2005a) distribution under the null of equal accuracy in population. The third and fourth terms (involving Γ_4 and Γ_5) arise due to weak predictability. The fourth term, Γ_5, corresponds to a non-centrality term that

gives some indication of the power that the test statistic has against deviations from the null hypothesis of equal population-level predictive ability $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}) = 0$ for all t — for which it must be the case that $\beta_w = 0$.

Under the assumptions of Clark and McCracken (2011a), it is straightforward to show that the mean of the asymptotic distribution of the MSE-F statistic can be used to approximate the mean difference in the average out-of-sample predictive ability of the two models, as:

$$E\sum_{t=R}^{T-\tau}(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) \approx \int_{\lambda}^{1}[-\omega^{-1}tr((-JB_1J' + B_2)V) + \beta'_w F_2^{-1}\beta_w]d\omega,$$

where $V = \lim_{T\to\infty} Var(T^{-1/2}\sum_{j=1}^{T-\tau} h_{2,j+\tau})$ for $h_{2,j+\tau}$ defined in section 3.1.2. Intuitively, one might consider using these expressions as a means of characterizing when the two models have equal average finite-sample predictive ability over the out-of-sample period. For example, having set these two expressions to zero, integrating and solving for the marginal signal-to-noise ratio implies $\beta'_w F_2^{-1}\beta_w / tr\left((-JB_1J' + B_2)\,V\right)$ equals $-\ln(\lambda)/(1-\lambda)$.[11] This condition simplifies further when $\tau = 1$ and the forecast errors are conditionally homoskedastic, in which case $tr((-JB_1J' + B_2)V) = \sigma^2 k_w$.

The marginal signal-to-noise ratio $\beta'_w F_2^{-1}\beta_w / tr\left((-JB_1J' + B_2)\,V\right)$ forms the basis of our new approach to testing for equal predictive ability. Rather than testing for equal population-level predictive ability $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}) = 0$ for all t — for which it must be the case that $\beta_w = 0$ — we test for equal average out-of-sample predictive ability $H_0 : \lim_{P,R\to\infty} E(\sum_{t=R}^{T-\tau}(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau})) = 0$ – for which it is the case that $\beta'_w F_2^{-1}\beta_w$ equals $\frac{-\ln(\lambda)}{1-\lambda} tr((-JB_1J' + B_2)V)$ for the recursive forecasting scheme and $tr((-JB_1J' + B_2)V)$ for the rolling scheme.

Since tabulating critical values in the general case is infeasible, Clark and McCracken (2011a) present a simple bootstrap that can provide asymptotically valid critical values in certain circumstances. In the following, let $B_i(T) = (T^{-1}\sum_{s=1}^{T-\tau} x_{i,s}x'_{i,s})^{-1}$ and $F_2(T) = J'_w B_2(T) J_w$, and let $V(T)$ denote a HAC estimator of the long-run variance of the OLS moment condition $\hat{v}_{2,s+\tau}x_{2,s}$ associated with the unrestricted model. The steps of the bootstrap are as follows.

1. (a) Estimate the parameter vector β_2^* associated with the unrestricted model using

[11] Under the rolling scheme the corresponding result is that $\beta'_w F_2^{-1}\beta_w / tr\left((-JB_1J' + B_2)\,V\right) = 1$.

the weighted ridge regression

$$\begin{aligned}\tilde{\beta}_{2,T} &= (\tilde{\beta}'_{1,T}, \tilde{\beta}'_{w,T})' \\ &= \arg\min_{b_2} \sum_{s=1}^{T-\tau} (y_{s+\tau} - x'_{2,s} b_2)^2 \text{ s.t. } b_2' J_w F_2^{-1}(T) J_w' b_2 = \hat{\rho}/T,\end{aligned} \tag{12}$$

where $\hat{\rho}$ equals $\frac{-\ln(\hat{\lambda})}{1-\hat{\lambda}} tr((-JB_1(T)J' + B_2(T))V(T))$ or $tr((-JB_1(T)J' + B_2(T))V(T))$ for the recursive or rolling schemes, respectively. Store the fitted values $x'_{2,t}\tilde{\beta}_{2,T}$. (b) Estimate the parameter vector β_2^* associated with the unrestricted model using OLS and store the residuals $\hat{v}_{2,s+\tau}$.

2. If $\tau > 1$, use NLLS to estimate an $MA(\tau - 1)$ model for the OLS residuals $\hat{v}_{2,s+\tau}$ such that $v_{2,s+\tau} = \varepsilon_{2,s+\tau} + \theta_1 \varepsilon_{2,s+\tau-1} + \ldots + \theta_{\tau-1}\varepsilon_{2,s+1}$.

3. Let η_s, $s = 1, \ldots, T$, denote an $i.i.d$ $N(0,1)$ sequence of simulated random variables. If $\tau = 1$, form a time series of innovations $\widehat{v}^*_{2,s+1} = \eta_{s+1}\widehat{v}_{2,s+1}$. If $\tau > 1$, form a time series of innovations computed as $\widehat{v}^*_{2,s+\tau} = (\eta_{s+\tau}\widehat{\varepsilon}_{2,s+\tau} + \widehat{\theta}_1\eta_{s-1+\tau}\widehat{\varepsilon}_{2,s+\tau-1} + \cdots + \widehat{\theta}_{\tau-1}\eta_{s+1}\widehat{\varepsilon}_{2,s+1})$, $s = 1, \ldots, T - \tau$.

4. Form artificial samples of $y^*_{s+\tau}$ using the fixed regressor structure, $y^*_{s+\tau} = x'_{2,s}\tilde{\beta}_{2,T} + \hat{v}^*_{2,s+\tau}$.

5. Using the artificial data, construct forecasts and an estimate of the test statistics (e.g., MSE-F, MSE-t) as if these were the original data.

6. Repeat steps 3-5 a large number of times: $j = 1, \ldots, N$.

7. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100 - \alpha)$%-ile of the empirical distribution of the simulated test statistics.

Clark and McCracken (2011a) show that critical values from this bootstrap are asymptotically valid in two important cases. First, if the number of additional predictors (k_w) is 1, then the bootstrap is asymptotically valid and allows for both multiple-step-ahead forecasts and conditionally heteroskedastic errors. Second, if the forecast horizon (τ) is 1 and the forecast errors are conditionally homoskedastic, then the bootstrap is asymptotically valid even when the number of additional predictors is greater than 1. While neither case covers the broadest situation in which β_w is not scalar and the forecast errors exhibit either serial correlation or conditional heteroskedasticity, these two special cases cover a wide range of empirically relevant applications. Kilian (1999) argues that conditional homoskedasticity is a reasonable assumption for one-step ahead forecasts of quarterly macroeconomic variables. Moreover, in many applications in which a nested model comparison is made (Goyal and

Welch (2008), Stock and Watson (2003), etc.), the unrestricted forecasts are made by simply adding one lag of a single predictor to the baseline restricted model. Of course, in more general settings that fall outside these two cases, it is possible that the proposed bootstrap will be reliable even if we can't prove its asymptotic validity. Some supplementary Monte Carlo experiments in Clark and McCracken (2011a) confirm this supposition on the broader reliability of our testing approach.

3.2.3 Calhoun (2011)

A completely different approach to managing the bias-variance trade-off is taken by Calhoun (2011). To understand his theoretical results, recall the logic behind the Giacomini and White (2006) approach. Their method was fundamentally built on the idea of preventing the parameter estimates from being consistent for the unknown true regression parameters. They achieved this by requiring that the parameters be estimated using a fixed rolling window of width R. Instead, Calhoun (2011) achieves the same lack of covergence by allowing the dimensionality of the model to increase with the sample size — while still allowing the initial sample size R to diverge to infinity. In doing so, he permits estimation error to contribute to the expected difference in squared forecast errors even in large samples. This approach to inference is in sharp contrast to the vast majority of the out-of-sample literature which assumes that the number of estimated parameters is fixed and finite.

In addition to allowing the dimensionality of the model to increase with the sample size, Calhoun (2011) is able to obtain an asymptotically normal test statistic by choosing a very specific null hypothesis — one that differs substantially from that addressed by either Giacomini and White (2006) or Clark and McCracken (2011a). This hypothesis is based on the idea that there exists hypothetical future forecast origins after time T and the testing procedure is intended to select the model that we expect will forecast most accurately in that hypothetical future. Specifically, suppose that at the end of our existing sample $s = 1, \ldots, T$, there exists Q future forecast origins $t = T, \ldots, T + Q - 1$ and define $\bar{D}_1$ and $\bar{D}_2$ as

$$\begin{aligned} \bar{D}_1 &= (P - \tau + 1)^{-1} \textstyle\sum_{t=R}^{T-\tau} (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) \\ \bar{D}_2 &= Q^{-1} \textstyle\sum_{t=T}^{T+Q-1} (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2), \end{aligned}$$

where $\hat{u}_{i,t+\tau}$ $i = 1, 2$ denote forecast errors associated with OLS estimated restricted and unrestricted models, respectively. The null hypothesis takes the form $H_0 : E(\bar{D}_2 | \Im_T) = 0$

where $E(.|\Im_T)$ is the conditional expectation operator given information available at time T. With this null hypothesis in hand, and assuming that $P \to \infty$, $R \to \infty$, $P^2/T \to 0$, $Q \to \infty$, and k_2/T is uniformly positive, he shows that

$$(P - \tau + 1)^{1/2}(\bar{D}_1 - E(\bar{D}_2|\Im_T)) \to^d N(0, S_{\hat{f}\hat{f}}), \tag{13}$$

where k_2 is the number of additional regressors in the unrestricted model and $S_{\hat{f}\hat{f}} = \lim_{P,R\to\infty} Var((P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau}((\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) - \bar{D}_1))$. Note that, as was the case for Giacomini and White (2006), the asymptotic variance differs from that in West (1996) since $S_{\hat{f}\hat{f}} \neq S_{ff}$.

3.3 Applications: pairs of models

To illustrate the use of some of the tests described in this section, we provide results from an application to forecasting U.S. inflation. Specifically, we forecast inflation in the GDP price index, using quarterly data. We consider simple autoregressive models and reduced-from Phillips curve models that augment the autoregressive specification to include indicators of economic activity, either GDP growth or the deviation of GDP from a statistical estimate of trend. In keeping with the theory of this section, we abstract from data revisions and simply measure all variables with the currently available vintage of data, taken from the FAME database of the Federal Reserve Board of Governors.

In light of evidence of significant shifts in trend inflation (see, e.g., Kozicki and Tinsley (2001), Stock and Watson (2007), and Clark and Doh (2011)), we use models of inflation relative to a lagged trend, where trend is defined as the long-run inflation forecast from a survey of professional forecasters. Our models take the forms

$$\pi^{(\tau)}_{t+\tau} - \pi^*_t = \alpha_0 + \sum_{l=1}^{L} \alpha_l(\pi^{(\tau)}_{t-l+1} - \pi^*_{t-l+1-\tau}) + e_{t+\tau} \tag{14}$$

$$\pi^{(\tau)}_{t+\tau} - \pi^*_t = \beta_0 + \sum_{l=1}^{L} \beta_l(\pi^{(\tau)}_{t-l+1} - \pi^*_{t-l+1-\tau}) + \beta_{L+1}x_t + u_{t+\tau}, \tag{15}$$

where $\pi^{(\tau)}_t = (400/\tau)\log(P_t/P_{t-\tau})$, π^*_t = trend inflation measured with the survey-based estimate PTR used in the Federal Reserve Board's FRB/US model, and x_t is either GDP growth computed as $100\log(GDP_t/GDP_{t-\tau})$ or the GDP gap computed as $100\log(GDP_t/GDP^*_t)$, using the 1-sided moving average filter of Hallman, Porter, and Small (1991) to estimate the trend GDP^*_t.[12]

[12]The FRB/US measure of (5- to 10-year-ahead) inflation expectations splices econometric estimates of

For models of this form, in this section we report results for forecast horizons of $\tau = 1$ and $\tau = 4$ quarters, for a sample of 1985:Q1 $+\tau - 1$ through 2011:Q4. In most cases, our results include both the recursive and rolling estimation schemes. In the recursive case, the models are estimated with data samples starting in 1962:Q2. In the rolling case, the size of the estimation sample is kept fixed at the size of the sample used in estimating coefficients for the first forecast, for period 1985:Q1 $+\tau - 1$. At the 1-quarter forecast horizon, we include two lags of inflation in the models; at the 4-quarter horizon, we use just one lag of inflation.

3.3.1 Non-nested models: population-level predictive ability

Table 3 presents results for tests of equal predictive accuracy at the population level applied to forecasts from non-nested models. Both models take the form of equation (15); the first uses GDP growth to measure economic activity, and the second uses the GDP gap. According to West (1996), a t-test for equal MSE can be computed without a need to correct for any effects of parameter estimation. Accordingly, the table reports a simple t-test computed as in Diebold and Mariano (1995), using the variance correction developed in Harvey, Leybourne, and Newbold (1997). At both forecast horizons and under both estimation schemes, the model with GDP growth yields an MSE lower than does the model with the GDP gap. However, at the 1-quarter horizon, differences in MSEs are small, and the t-statistics indicate the null of equal accuracy cannot be rejected. At the 4-quarter horizon, the differences in MSEs are larger, but the null of equal accuracy still cannot be rejected at the 5 percent level. At the 10 percent level, the null of equal accuracy can be rejected under the rolling scheme, but not the recursive.

3.3.2 Nested models: population-level and finite-sample predictive ability

Table 4 presents results for tests of equal predictive accuracy applied to forecasts from nested models, at both the population level and in the finite sample. The null model is the autoregression given in equation (14); the alternative is the Phillips curve of equation (15). We consider two different nested model comparisons, one with GDP growth in the Phillips curve and the other with the GDP gap in the Phillips curve. The table provides RMSEs

inflation expectations from Kozicki and Tinsley (2001) early in the sample to 5- to 10-year-ahead survey measures compiled by Richard Hoey and, later in the sample, to 10-year-ahead values from the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. The estimate of the output gap obtained with the filter of Hallman, Porter, and Small (1991) is highly correlated with an output gap based on the Congressional Budget Office's estimate of potential output.

relative to the AR benchmark and the MSE-t and MSE-F tests for equal accuracy, for both alternative models, both horizons, and both estimation schemes.[13] As the RMSE ratios indicate, the Phillips curve with GDP growth is consistently more accurate than the AR benchmark, while the Phillips curve with the GDP gap is consistently less accurate than the AR benchmark.

We begin with testing the null of equal forecast accuracy at the population level. As noted above, in this case, because the models are nested, the MSE-F and MSE-t tests have non-standard distributions; as a consequence, one cannot simply (and correctly) apply a Diebold-Mariano-West test with standard normal critical values. A researcher who did so would be very unlikely to reject the null hypothesis of equal accuracy at the population level (Monte Carlo evidence in, e.g., Clark and McCracken (2001, 2005) show such a test to be severely undersized). Indeed, in this application, the p-values of MSE-t tests compared against standard normal critical values are all above 18 percent.

Correctly testing the null of equal forecast accuracy at the population level requires using asymptotic critical values tabulated in Clark and McCracken (2001) and McCracken (2007), simulated via Monte Carlo as in Clark and McCracken (2005), or bootstrapped. In this application, we use p-values computed with the fixed regressor bootstrap described in section 3.1.4 above, denoted "FRBS, population EPA" in the table. In the case of the Phillips curve with GDP growth versus the AR benchmark, the p-values for the MSE-F test are all below 5 percent, rejecting the null of equal accuracy at the population level. Consistent with evidence in Clark and McCracken (2011b) on the relative power of the MSE-t and MSE-F tests, the p-values are higher for the MSE-t test. For GDP growth, the MSE-t test rejects the null of equal accuracy at the 10 percent level for the 1-quarter horizon, but the null cannot be rejected at the 4-quarter horizon. In the case of the Phillips curve with the GDP gap versus the AR benchmark, none of the MSE-F or MSE-t tests reject the null of equal accuracy at the population level.

Now consider testing equal accuracy in the finite sample. The finite sample hurdle is higher: one model may be more accurate than another in population but not in the finite sample, due to imprecision in estimating parameters in the finite sample. To evaluate the finite sample null hypothesis, a researcher has two basic choices. The first is to rely on the asymptotics of Giacomini and White (2006), which require either a fixed or rolling

[13] In all cases, we compute the MSE-t test using the the variance correction developed in Harvey, Leybourne, and Newbold (1997).

estimation scheme, but yield a standard normal distribution for the MSE-t test.[14] An alternative approach is to rely on the asymptotics of Clark and McCracken (2011a), which apply under any estimation scheme, but yield non-standard asymptotic distributions for the MSE-F and MSE-t tests and thereby require the use of a bootstrap to obtain critical values.[15] In this application, we compute p-values with a fixed regressor bootstrap under the null of equal forecast accuracy in the finite sample, denoted "FRBS, finite-sample EPA" in Table 4.

As expected, the evidence that a reduced-form Phillips Curve forecasts better than an AR model is weaker for the null of equal finite-sample predictive ability than for the null of equal population-level predictive ability. Using the (standard-normal) asymptotics of Giacomini and White (2006), none of the MSE-t tests reject the null of equal accuracy in the finite sample. For the model with GDP growth, the p-values (across horizons) are all on the order of 0.20; for the model with the GDP gap, the p-values are all above 0.50. Similarly, under the approach of Clark and McCracken (2011a), neither the MSE-t nor (more powerful) MSE-F test rejects the null of equal accuracy in the finite sample, at conventional significance levels.

3.3.3 Overlapping models

Table 5 presents results for tests of equal predictive accuracy applied to forecasts from potentially overlapping models, under a recursive estimation scheme. Both of the models take the Phillips Curve form of equation (15); the first uses GDP growth to measure economic activity, and the second uses the GDP gap. Following Clark and McCracken's (2012) two-step testing procedure for testing equal accuracy, we first compare the test based on the variance of the loss differential to critical values obtained with the fixed regressor bootstrap. (Note that we report the simple variance $\hat{S}_{dd}$ rather than the scaled version $P\hat{S}_{dd}$ that has a non-degenerate asymptotic distribution; ignoring the scaling has no effect on the inferences drawn under the bootstrap.) The p-values of these tests are very low, clearly rejecting the null of overlapping models. Having rejected at the first stage of the test, we proceed to

[14] While the results of Giacomini and White (2006) do not apply under a recursive estimation scheme, in this application we provide normal distribution-based p-values for this scheme anyway, since Monte Carlo evidence in Clark and McCracken (2011a) suggests that, in practice, comparing the MSE-t test against standard normal critical values works about as well for the recursive scheme as for the rolling.

[15] As noted above, the asymptotic validity of the bootstrap of Clark and McCracken (2011a) requires that either the forecasts be serially uncorrelated and conditionally homoskedastic or that the number of additional regressors in the larger model is equal to 1.

the second stage, of comparing the MSE-t statistic against standard normal critical values, as appropriate with non-nested models. With the MSE-t test taking the value of -1.125 at the one-quarter horizon and -1.428 at the four-quarter horizon, it falls short of normal critical values at a 10 percent confidence level, so we cannot reject the null of equal accuracy between the non-nested models (as in the non-nested test results also presented in Table 3).

In this application, using the conservative one-step procedure of Clark and McCracken (2012) yields the same conclusion. Under this approach, we compare the t-test for equal MSE to the critical values shown in the last two rows of the table. The lower tail critical value is the minimum of the lower tail critical values from the standard normal and bootstrap distributions; the upper tail critical value is the maximum of the upper tail critical values from the standard normal and bootstrap distributions. With the MSE-t test statistic not close to these critical values, we cannot reject the null of equal forecast accuracy.

3.3.4 Summary of application results

Putting together all of this section's application results, it seems that there is some evidence to indicate that, at the population level, GDP growth is helpful for forecasting inflation (based on tests against an AR benchmark). There is less evidence to suggest that a GDP gap is helpful, but at the population level, the differences in accuracy for a model with GDP growth versus a model with the GDP gap aren't large enough to be statistically significant. It does not appear to be the case that the true model is an AR specification: the null hypothesis that the competing Phillips Curves with GDP growth and the GDP gap are overlapping is soundly rejected.

In contrast, at the finite sample level, there is no evidence of statistically significant predictive ability. It seems that, with limited data samples, the parameters of the Phillips curve are estimated imprecisely enough that an AR model, while not the true model, is about as accurate as a Phillips curve.

4 Unconditional Versus Conditional Evaluation

In section 3 we introduced the distinction between tests of population-level predictive ability and tests of finite-sample predictive ability. There, the key distinction was the importance of introducing finite sample estimation error under the null hypothesis. That is, tests of population-level predictive ability test the null hypothesis that $Ef_{t+\tau}(\beta^*) = \gamma$, whereas

tests of finite-sample level predictive ability test the related but distinct hypothesis akin to $Ef_{t+\tau}(\hat{\beta}_t) = \gamma$.

One thing both hypotheses have in common is that the expectation operator $E(\cdot)$ is defined relative to the trivial σ-field $(\varnothing, \Im)$ and hence is an unconditional expectation. In the terminology of this section, everything that has been discussed so far in this chapter can be characterized as a test of unconditional predictive ability.[16] In contrast, Giacomini and White (2006) consider a different type of hypothesis in which they replace the unconditional expectation operator with a conditional one $E[\cdot|\Im_t]$, where $\Im_t$ denotes an information set available to the forecasting agent at time t. This somewhat subtle difference leads to a broader class of tests of predictive ability.

As an example of how such a test might be useful, consider a proposal suggested, but not elucidated, in Diebold and Mariano (1995). They suggest that while it might be the case that two non-nested models have equal (unconditional) predictive ability in terms of mean square errors, it still might be the case that one model performs better than the other at certain parts of the business cycle and vice versa.[17] To see how this might occur, first consider constructing a test of equal unconditional MSE via a regression of the form

$$u^2_{1,t+\tau} - u^2_{2,t+\tau} = \alpha_0 + \varepsilon_{t+\tau}. \tag{16}$$

In this notation the null hypothesis $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}) = 0$ simplifies to testing the null $H_0 : \alpha_0 = 0$. Now suppose that instead of estimating the regression in equation (16) we estimate one of the form

$$u^2_{1,t+\tau} - u^2_{2,t+\tau} = \alpha_0 + \alpha_1 1(\aleph_t) + \varepsilon_{t+\tau}, \tag{17}$$

where $1(\cdot)$ denotes a function taking the value 1 if the argument is true and zero otherwise and $\aleph_t$ denotes the event that the economy is in a recession at time t. In this notation, the null hypothesis $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}) = 0$ is equivalent to testing the null $H_0 : \alpha_0 + \alpha_1 d = 0$, where d denotes the percentage of the sample that the economy is in a recession.

While the regression in equation (17) is unnecessarily complicated for testing the null of equal unconditional predictive ability, it opens the door for tests of the kind that Diebold and Mariano (1995) proposed. For example we could use the regression in (17) to test the null that the two models have equal predictive ability regardless of the state of the business

[16] Calhoun (2011) being an important caveat.

[17] We'll return to the issue of nested models later in this section.

cycle — that is, $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph_t) = E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph'_t) = 0$ — by testing whether $H_0 : \alpha_0 = \alpha_1 = 0$ holds. If this more restrictive hypothesis holds then it is certainly the case that the weaker hypothesis of equal predictive ability over the entire business cycle holds but the converse is not true. One could have $\alpha_0 + \alpha_1 d = 0$ and yet both α_0 and α_1 are not zero and hence it is possible that one model forecasts better than the other depending on the state of the business cycle.

In this discussion, we have purposefully shied away from the population versus finite-sample predictive ability issue. We did so in order to emphasize that the concept of conditional predictive ability is a completely distinct concept. Tests of conditional predictive ability can be implemented at both the finite-sample and population level. To see how, consider the slightly modified version of the regression in equation (17):

$$\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau} = \alpha_0 + \alpha_1 1(\aleph_t) + \varepsilon_{t+\tau}. \tag{18}$$

The sole modification is that we wrote the regression in terms of the estimated forecast errors $\hat{u}^2_{i,t+\tau}$ rather than the population values of the forecast errors $u^2_{i,t+\tau}$. Whether we are testing for equal population level predictive ability regardless of the state of the business cycle [$H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph_t) = E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph'_t) = 0$] or equal finite-sample predictive ability regardless of the state of the business cycle [$H_0 : E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\aleph_t) = E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\aleph'_t) = 0$], this type of regression can be used as a testing device. What distinguishes the two is largely a matter of asymptotics. In the following we consider two alternative approaches. We conclude the section with an application to inflation forecasting.

4.1 Giacomini and White (2006)

While Diebold and Mariano (1995) first suggested the idea of conditional predictive ability, Giacomini and White (2006) first provided a theory for implementing such a test with forecasts that may come from estimated models and made the idea of conditional predictive ability a major part of the literature. Continuing with the recession-oriented example above, they suggest constructing a test statistic of the form[18]

$$GW_T = (P - \tau + 1)\bar{Z}'_T \hat{S}^{-1}_{\hat{f}\hat{f}} \bar{Z}_T, \tag{19}$$

[18] Under certain conditions, this test statistic is asymptotically equivalent to using the uncentered R^2 from the regression in equation (18) as the test statistic. These conditions are delineated in Giacomini and White (2006). For brevity we emphasize the more generally valid quadratic form in equation (19).

where $\bar{Z}_T$ denotes the vector

$$\left(\frac{1}{P-\tau+1} \sum_{t=R}^{T-\tau} (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2), \ \frac{1}{P-\tau+1} \sum_{t=R}^{T-\tau} (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) 1(\aleph_t) \right)' \tag{20}$$

and $\hat{S}_{\hat{f}\hat{f}}$ denotes an appropriately constructed covariance matrix associated with the asymptotic distribution of $(P-\tau+1)^{1/2}\bar{Z}_T$. Under modest mixing and moment conditions they show that the statistic GW_T is asymptotically χ^2 with 2 degrees of freedom.

In order for their asymptotics to work they make one additional assumption: that the models used to construct the forecasts are estimated using a rolling (or fixed) window of observations of size R that is finite and small relative to the prediction sample P. While this assumption rules out the use of the recursive scheme it has many powerful benefits, which we delineate below.

1. The GW_T statistic tests for not only conditional predictive ability but also finite-sample predictive ability. That is, the null hypothesis being tested is one that takes the form $H_0 : E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2 | \aleph_t) = E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2 | \aleph_t') = 0$. The reason for this, previously delineated in section 3.2.1, is that regardless of the total sample size T, estimation error never vanishes and hence estimation error is introduced under the null.

2. The test allows for both nested and non-nested comparisons in the same asymptotic framework. Regardless of whether models 1 and 2 are nested or non-nested, the GW_T statistic remains asymptotically χ^2 with 2 degrees of freedom

3. The statistic is far more generally applicable than testing the conditional predictive ability of two forecasting models over the business cycle. One could imagine testing for conditional zero-mean prediction error, efficiency, encompassing, etc. In each case there is some proposed null hypothesis of the form $H_0 : E(f_{t+\tau}(\hat{\beta}_t) | \Im_t) = 0$ where $\Im_t$ denotes an information set available to the forecasting agent at time t. If we let z_t denote a $k_z \times 1$ vector of instruments that is observable at time t, the null can be tested using the same statistic $GW_T = (P-\tau+1)\bar{Z}_T' \hat{S}_{\hat{f}\hat{f}}^{-1} \bar{Z}_T$ but where $\bar{Z}_T$ denotes the vector $(P-\tau+1)^{-1} \sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t) z_t$ and $\hat{S}_{\hat{f}\hat{f}}$ denotes a consistent estimate of the long-run variance of $(P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t) z_t$. In each application the statistic is asymptotically χ^2 with k_z degrees of freedom.

4. In constructing the test statistic it is important to ensure that the estimate $\hat{S}_{\hat{f}\hat{f}}$ of the long-run variance $S_{\hat{f}\hat{f}} = \lim_{P\to\infty} Var((P-\tau+1)^{1/2}\bar{Z}_T)$ is appropriately constructed. In particular we have to account for the fact that the null hypothesis of conditional predictive

ability imposes restrictions on not only the first moment of $f_{t+\tau}(\hat{\beta}_t)z_t$, but also the second moments. Under the null $E(f_{t+\tau}(\hat{\beta}_t)|\Im_t) = 0$, $f_{t+\tau}(\hat{\beta}_t)z_t$ has an $MA(\tau-1)$ serial correlation structure. In contrast, a test of unconditional predictive ability only imposes restrictions on the first moment of $f_{t+\tau}(\hat{\beta}_t)z_t$.

To insure clarity of the point being made with regard to the asymptotic variance matrix $S_{\hat{f}\hat{f}}$, consider the simplest situation where $z_t = 1$. Under the null of equal finite-sample *unconditional* predictive ability we know from section 3.2.1 that

$$(P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) \rightarrow^d N(0, S_{\hat{f}\hat{f}}) \tag{21}$$

where $S_{\hat{f}\hat{f}} = \lim_{P\rightarrow\infty} Var((P-\tau+1)^{-1/2} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}))$. For this null hypothesis the structure of $S_{\hat{f}\hat{f}}$ is unconstrained in the sense that $\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$ may exhibit serial correlation of any order – including infinite. Hence one typically would estimate $S_{\hat{f}\hat{f}}$ as in Newey and West's (1987) HAC estimator by weighting the relevant leads and lags of the estimated covariance matrices $\hat{\Gamma}_{\hat{f}\hat{f}}(j) = (P - \tau + 1)^{-1} \sum_{t=R+j}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau})(\hat{u}^2_{1,t+\tau-j} - \hat{u}^2_{2,t+\tau-j})$, where $\hat{\Gamma}_{\hat{f}\hat{f}}(j) = \hat{\Gamma}_{\hat{f}\hat{f}}(-j)$.

Let's now return to the case where we want to test for equal finite-sample *conditional* predictive ability. We still obtain the result that

$$(P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) \rightarrow^d N(0, S_{\hat{f}\hat{f}}), \tag{22}$$

but the value of $S_{\hat{f}\hat{f}}$ *is now different.* In the notation above, due to the conditioning we know that for all $\tau \leq j$, $\hat{\Gamma}_{\hat{f}\hat{f}}(j) = 0$. Hence an asymptotically valid estimate of $S_{\hat{f}\hat{f}}$ now only requires estimating $\hat{\Gamma}_{\hat{f}\hat{f}}(j)$ for $0 \leq j \leq \tau - 1$. Despite these added restrictions, one certainly could continue to use a HAC estimator such as Newey and West's (1987), but that is likely to be unnecessarily profligate in the number of estimated covariances and may lead to size distortions of the kind discussed in section 7. A more parsimonious approach is simply to use a rectangular kernel that weights equally only the first $\tau - 1$ covariances.

4.2 West (1996)

In the Giacomini and White (2006) framework described above, by default one tests for both finite-sample predictive ability and conditional predictive ability and hence the null is of the form $E(f_{t+\tau}(\hat{\beta}_t)|\Im_t) = 0$. This occurs due to the nature of the small rolling (or fixed) window being used for estimating the model parameters. If instead we wanted to test

for conditional population-level predictive ability $E(f_{t+\tau}(\beta^*)|\Im_t) = 0$, we could do so using an appropriately modified version of the theory described in West (1996) that accounts for the fact that under the null hypothesis, $f_{t+\tau}(\beta^*)$ is unpredictable using any observables contained in the information set $\Im_t$.

As an example, let's revisit the recession example above where we are considering the relative predictive ability of two non-nested models. In the notation of West (1996), the null hypothesis of interest is $H_0 : E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph_t) = E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\aleph'_t) = 0$, where $u_{i,t+\tau}$, $i = 1, 2$, denote the population-level forecast errors associated with models 1 and 2, respectively. To test such a hypothesis it is reasonable to follow the intuition in Giacomini and White (2006) and base inference on the sample moment condition $\bar{Z}_T$ equal to

$$\left((P-\tau+1)^{-1} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}),\ (P-\tau+1)^{-1} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) 1(\aleph_t) \right)' \tag{23}$$

with corresponding test statistic

$$(P-\tau+1)\bar{Z}'_T \hat{\Omega}^{-1} \bar{Z}_T. \tag{24}$$

Interestingly, one is still able to use the asymptotic theory in West (1996) to show that this statistic can be asymptotically χ^2 with 2 degrees of freedom despite the fact that tests of conditional predictive ability are not discussed in that paper.

To see how, suppose that instead of wanting to test for conditional predictive ability, one wanted to test the null that the bivariate *unconditional* moment condition $(E(u^2_{1,t+\tau} - u^2_{2,t+\tau}),\ E(u^2_{1,t+\tau} - u^2_{2,t+\tau})1(\aleph_t))'$ is equal to zero. The results in West (1996) apply directly and we conclude that $(P-\tau+1)\bar{Z}'_T \hat{\Omega}^{-1} \bar{Z}_T \rightarrow^d \chi^2(2)$ for an appropriately estimated (2×2) variance matrix Ω. Now suppose that instead we impose the *strictly stronger* conditional moment condition $E(u^2_{1,t+\tau} - u^2_{2,t+\tau}|\Im_t) = 0$. It must still be the case that $(P-\tau+1)\bar{Z}'_T \hat{\Omega}^{-1} \bar{Z}_T \rightarrow^d \chi^2(2)$ for an appropriately estimated variance matrix Ω.

The main difference between the two cases just described, as we noted above for the Giacomini and White (2006) analytics, is that a null of conditional predictive ability imposes a restriction on both the first *and* second moments of $f_{t+\tau}(\beta^*)z_t$. In particular $f_{t+\tau}(\beta^*)z_t$ has an $MA(\tau-1)$ serial correlation structure. This changes how we estimate the asymptotic variance Ω via how we estimate both the S_{ff} and S_{fh} components in equation (2). Specifically, both of these two matrices can now be estimated using a HAC estimator with a rectangular kernel of order $\tau - 1$, whereas when testing for unconditional predictive ability

one would have had to account for the possibility that $f_{t+\tau}(\beta^*)z_t$ exhibited serial correlation of infinite order using a HAC estimator such as that of Newey and West (1987).

4.3 Application

To illustrate testing of conditional equal predictive ability, we take up an example like the one described early in this section. For the two Phillips Curve models based on GDP growth and the GDP gap, we test whether the models have equal predictive ability regardless of the state of the business cycle. For that purpose, we form the Giacomini and White (2006) test statistic given above in equations (19) and (20), using the NBER's definition of recession periods. We compute the test statistics with a variance obtained from a rectangular kernel and a bandwidth equal to the forecast horizon less 1. We report in Table 3 both the test statistic and the p-value from the χ^2 distribution with two degrees of freedom. While the results of Giacomini and White (2006) do not apply under a recursive estimation scheme, we provide p-values for tests obtained under this scheme anyway, in part because, as explained above, West's (1996) asymptotics can also justify the test, although his results actually require a variance estimate different from the one we use. The results of Giacomini and White's conditional test indicate that, at both horizons and under both schemes, the null of equal predictive ability regardless of the state of the business cycle cannot be rejected. The p-values of the test are all above 0.20. Combined, with the other, previously discussed results in Table 3 for this non-nested application, neither the null of unconditional predictive ability nor the null of conditional predictive ability can be rejected.

5 Evaluation of Multiple Forecasts

In each of the previous sections we focused on tests of equal predictive ability between two models. In practice, however, it is sometimes the case that there are several, or perhaps even many, models being compared. As summarized in West (2006), early work on the evaluation of multiple forecasts focused on one-step testing procedures for non-nested models or judgmental forecasts. In recent years, there has been significant progress on methods for evaluating multiple forecasts, including extensions to nested model comparisons and stepwise procedures.[19] In the following we provide a brief summary of the existing procedures

[19] See Corradi and Swanson (2012) for a recent survey of methods for evaluating multiple forecasts. Corradi and Swanson (2012) also propose a new test linked to the concept of stochastic dominance, which can be seen as combining reality check testing with forecast combination.

as of West's review as well as discuss more recent innovations.

5.1 One-step procedures

White (2000) was the first to develop a procedure for testing equal predictive ability among many models. His test statistic takes the form

$$\max_{k=1,\ldots,K} \bar{d}_k$$

where $\bar{d}_k = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{d}_{k,t+\tau}$, the out-of-sample average loss differential between model $k = 1,\ldots,K$ and the benchmark model 0. When the loss function is quadratic the loss differential takes the form $\hat{d}_{k,t+\tau} = \hat{u}^2_{0,t+\tau} - \hat{u}^2_{k,t+\tau}$ but the theory allows for a much wider class of loss functions. Under the null that each of the competing models is equally accurate, the results of West (1996) in section 3.1.1 imply that for each k, $P^{1/2}\bar{d}_k$ is asymptotically normal and hence $\max_{k=1,\ldots,K} \bar{d}_k$ converges in distribution to the maximum of K correlated zero mean normal random variates.

The main innovation in White is the introduction of a bootstrap-based approach to constructing asymptotically valid critical values associated with the asymptotic distribution of $\max_{k=1,\ldots,K} \bar{d}_k$. His "Reality Check" bootstrap has the advantage of only requiring the resampling of forecast errors, thus avoiding the estimation of forecasting models and constructing forecasts using bootstrapped artificial data. One disadvantage of his bootstrap is that it is only applicable to judgmental forecast comparisions or to non-nested model comparisons with the additional assumption that the number of forecasts $P-\tau+1$ is small relative to the intitial estimation sample R.

In White (2000), and in most subsequent studies in the multiple model literature, the null and alternative hypotheses take the form

$$H_0 : \max_{k=1,\ldots,K} Ed_{k,t+\tau} \leq 0 \text{ vs. } H_A : \max_{k=1,\ldots,K} Ed_{k,t+\tau} > 0.$$

Note that if the null hypothesis is rejected, we only conclude that there exists at least one model that is more accurate than the baseline model. It is in this sense that White's testing procedure is "one-step." There exists no second step that allows one to identify the complete subset of models that are more accurate than the baseline model. We will return to this issue in the following subsection.

At the time of West's survey, two other studies had developed extensions of White's (2000) testing approach. First, Hansen (2005) shows that normalizing and re-centering the

test statistic in a specific manner can lead to a more accurately sized and powerful test, with the power-enhancing adjustments serving to reduce the influence of bad forecasting models. Second, under basic West (1996) asymptotics, Corradi and Swanson (2007) develop a bootstrap applicable when parameter estimation error is not irrelevant. Under general conditions, as forecasting moves forward in time and the model estimation window expands, observations earlier in the data sample enter in the forecast test statistics more frequently than do observations that fall later in the data sample. This creates a location bias in the bootstrap distribution. To adjust for this asymptotic bias, Corradi and Swanson develop a recentering of the bootstrap score. Under their West-type asymptotics, the bootstrap can be applied to forecasts from non-nested models.

More recently, Song (2012) builds on Hansen (2005) to further robustify the local asymptotic power of one-sided sup tests of predictive ability. Song proposes a hybrid test that couples the one-sided sup test with another test that has better power against some alternatives against which the one-sided sup test has weak power. This other, complementary test draws on Linton, Massoumi, and Whang's (2005) approach to testing stochastic dominance. Letting $\hat{d}(m)$ denote the loss differential for model i relative to the benchmark and M denote the total number of alternative models, the complementary test statistic is

$$T^S = \min\{\max_{m \in M} \hat{d}(m), \max_{m \in M}(-\hat{d}(m))\}.$$

Song develops the test as applying to nested models under under the asymptotics of Giacomini and White (2006), which make parameter estimation error irrelevant either through the use of a fixed or rolling estimation scheme or a forecast sample that is very small relative to the estimation sample.

Also building on Hansen (2005), Corradi and Distaso (2011) develop a class of tests for superior predictive ability, intended to have power better than White's (2000) reality check. Corradi and Distaso assume that parameter estimation error is asymptotically irrelevant, for reasons such as those given in West (1996) — for example, a forecast sample that is small relative to the estimation sample. Drawing on the literature on constructing confidence intervals for moment conditions defined by multiple inequalities, Corradi and Distaso develop a class of tests for superior predictive ability, which can be compared to bootstrapped critical values. Their general class of tests includes Hansen's (2005) SPA test — the maximum across models of t-tests for equal loss (e.g, equal MSE).

Other recent extensions to the literature on evaluating multiple forecasts have focused

on projections from nested models.[20] To evaluate forecasts from a small to modest set of nested models, Rapach and Wohar (2006) rely on an expanded version of the restricted VAR bootstrap used by such studies as Kilian (1999) and Clark and McCracken (2005a) to evaluate pairs of forecasts. This approach consists of comparing the maximum of forecast test statistics (e.g., MSE-F and ENC-F) to a bootstrapped distribution obtained by: simulating data from a VAR in the predictand of interest and all predictors considered, where the equation for the predictand y is restricted to the form of the null model; and then generating forecasts and test statistics for all models considered.

Motivated in part by a desire to avoid the computations associated with these kinds of bootstrap methods, Hubrich and West (2010) propose taking advantage of the approximate normality (or exact normality with rolling forecasts and a null model that is a martingale difference sequence) of the Clark and West (2006, 2007) test (equivalently, the ENC-t test). One test statistic they propose is a χ^2 test. Letting $\overline{CW}$ denote the mean of the vector of numerators of the Clark and West-t test (loss differentials) and $\hat{S}_{CW,CW}$ denote the estimated (long-run) variance-covariance matrix of the vector of loss differentials, the test statistic is formed as $(P-\tau+1)\overline{CW}'\hat{S}_{CW,CW}^{-1}\overline{CW}$. The other test statistic they propose is the maximum of the sequence of Clark and West t-tests for all models considered. These tests can be viewed as tests of equality of adjusted MSEs from multiple forecasts or multiple encompassing tests.

Taking the individual t-tests to be normally distributed, the quantiles of the maximum distribution can either be easily computed with simple Monte Carlo simulations or, when the model set is very small, looked up in Monte-Carlo generated tables provided by Hubrich and West.[21] In general settings, using the Hubrich-West result involves computing a variance-covariance matrix for the vector of loss differentials for the set of models, conducting Monte Carlo simulations of a multivariate normal distribution with that variance-covariance matrix, and computing quantiles of the simulated distribution of the maximum statistic.

Granziera, Hubrich, and Moon (2011) propose a likelihood ratio-type predictability test

[20] In another recent extension for non-nested models, Mariano and Preve (2012) propose a multivariate version of the Diebold and Mariano (1995) test for application to forecasts that either do not come from estimated models or, if they do, come from models estimated with samples large enough relative to the forecast sample as to make parameter estimation error irrelevant. Under their assumptions, a Wald-type test in the vector of loss differentials has a χ^2 distribution.

[21] For the case of three forecasts (which yields two loss differentials), Hubrich and West (2010) provide tables of critical values obtained by numerical solution of the density function of the maximum of two correlated standard normal random variables. The appropriate critical value is a function of the correlation between the loss differentials.

for comparison of a small set of nested models. Their proposed test distinguishes among different types of nesting relationships, with all alternative models nesting the benchmark specification: (1) all of the alternative models nest another, (2) no nesting relationship among the alternative models, and (3) nesting within certain groups of models but not across groups. By adjusting the (two-sided) Wald statistic of Hubrich and West (2010) to formulate it as one-sided test, Granziera, Hubrich, and Moon improve the power of the test. Under asymptotics similar to studies such as Clark and McCracken (2001, 2005a), Granziera, Hubrich, and Moon show the limiting distribution of their proposed test to be a functional of Brownian motion. Following Hubrich and West (2010) in treating the underlying loss differentials — numerators of the Clark and West (2006, 2007) test — as approximately normally distributed, Granziera, Hubrich, and Moon propose comparing the likelihood ratio-type predictability test to χ^2 critical values. In light of the asymptotic results of Clark and McCracken (2001, 2005a) that indicate the t-test distribution for each forecast pair (alternative versus benchmark) is not actually normal under general conditions, Granziera, Hubrich, and Moon also compare their proposed test to critical values obtained with the bootstrap of Clark and McCracken (2011b).

Finally, under the large R, large P asymptotics of such studies as Clark and McCracken (2001, 2005a) and West (1996), Clark and McCracken (2011b) develop a fixed regressor bootstrap for testing population-level equal accuracy of forecasts from nested models. They define test statistics that are the maxima (across models) of the equal MSE and encompassing tests defined in section 3.2.2, where each of a range of alternative models is tested against a nested benchmark model. They show that the asymptotic distributions are the maxima of pairwise asymptotic distributions of MSE-F, MSE-t, ENC-F, and ENC-t tests that are functions of stochastic integrals of Brownian motion. Clark and McCracken develop a fixed regressor bootstrap for obtaining asymptotic critical values and prove the validity of the bootstrap, for a null hypothesis of equal population-level accuracy. The bootstrap takes the basic form given above in section 3.1.3, modified to account for multiple alternative models and to sample the needed residuals from an unrestricted model that includes all predictors considered across all models.

5.2 Stepwise Procedures

As noted above, a weakness of the White (2000), Hansen (2005), and other one-shot testing procedures is that they only inform the user whether or not their exists a competing model

that is more accurate than the baseline model. No additional information is given that tells the user if there is more than one model that is more accurate. As a practical matter it might be useful to know which models are more accurate. For example, suppose that the competing "models" are various trading rules and the baseline is a no-change position in a particular asset market. While it is certainly useful to know that one of the models is better than the baseline for forming asset allocation decisions, it would be even better to be able to identify all of the trading rules that are better so that an investor can diversify their portfolio across these trading rules as a hedge against market risk.

A step-wise multiple testing procedure is a straightforward extension of the one-shot procedures discussed above that allows one to identify the collection of models that are superior to the baseline model. The first to introduce such a procedure is Romano and Wolf (2005). There they note that if the one-shot procedure of White (2000) is iterated in a particular way then one can identify the collection of models that are more accurate than the benchmark model 0. The basic procedure is delineated below.[22]

Step 1: Relabel the out-of-sample average loss differentials $\bar{d}_j$ from smallest to largest so that $\tilde{d}_1 = \min_{k=1,\ldots,K} \bar{d}_k$ and $\tilde{d}_K = \max_{k=1,\ldots,K} \bar{d}_k$.

Step 2: Use the bootstrap procedure of White (2000) to estimate the critical value c_α associated with an α-level one-shot test as described above.

Step 3: If we fail to reject the null the procedure stops. If we reject, remove those models that have values of $\tilde{d}_j$ greater than c_α.

Step 4: Repeat Steps 1-3 but only using those models that have not been removed. The algorithm stops when no additional models are removed.

Romano and Wolf (2005) show that asymptotically this algorithm will identify all models that are more accurate than the benchmark in the sense that if model k is removed we know that $Ed_{k,t+\tau} > 0$. In addition, based primarily on the union-intersection principle, they show that in large samples the familywise error rate (FWE) will be bounded from above by α. Moreover, the FWE will equal α if there exists at least one model for which $Ed_{k,t+\tau} = 0$ and there exist no models for which $Ed_{k,t+\tau} < 0$.

One very useful aspect of the Romano and Wolf stepwise procedure is that the proofs do not rely specifically upon White's Reality Check bootstrap or the particular test statistic used by White. As such, one-shot extensions of White's work, like that made by Hansen

[22] Romano and Wolf (2005) describe several versions of this algorithm. For brevity we only describe their "Basic StepM Method" modified to our existing notation.

(2005), are readily adapted to the stepwise algrothm. In fact Hsu, Hsu, and Kuan (2010) do exactly that and provide analytical and Monte Carlo evidence that their procedure is particularly powerful at identifying those models that are more accurate than the benchmark while still controlling the FWE.

5.3 The Model Confidence Set

The multiple-forecast evaluation procedures described above are all tests for superior predictive ability: they determine whether a chosen benchmark forecast is significantly outperformed by any of the alternative forecasts considered. As such, the null hypothesis underlying tests for superior predictive ability is a composite hypothesis, involving multiple inequality conditions.

Hansen, Lunde, and Nason (2011) develop an alternative approach to inference with multiple models.[23] Their approach seeks to identify the subset of models that contains the best model with a given level of confidence without having to specify a particular model as a benchmark. Their procedure can be seen as yielding a confidence interval around the true model set in the same way that common, classical test statistics yield a confidence interval containing the true parameter value with probability no smaller than 1 less the chosen significance or confidence level. Unlike the tests for superior predictive ability, the model confidence set approach consists of a sequence of tests involving just equalities, thereby avoiding composite testing.

The asymptotic derivations of Hansen, Lunde, and Nason (2011) treat the number of models as fixed and the time series dimension as limiting to infinity. For simplicity, they abstract from the complexities that can arise with nested models. Their theory takes forecast loss differentials as primitives and assumes positive variances and stationarity of the differentials — conditions that will be violated under the population-level asymptotics of such nested model studies as Clark and McCracken (2011b). Hansen, Lunde, and Nason (2011) note that, for nested models, application of their methods requires either the Giacomini and White (2006) assumption that a fixed or rolling scheme to generating forecasts makes parameter estimation error irrelevant or some adjustment of the test statistics and bootstrap algorithm.

Determining the model confidence set involves a test for model equivalence and an elim-

[23] While we focus on the use of their methods for forecast evaluation, the results of Hansen, Lunde, and Nason (2011) include in-sample tests applied to regression models.

ination rule. In practice, this involves a set of t-statistics for equality of loss differentials, one for all (unique) combinations of loss differentials and another for the average loss differential of each model relative to a given benchmark. These are then used to form maximum t-statistics that serve as the basis for inference. Since the asymptotic distributions of these test statistics depend on nuisance parameters, Hansen, Lunde, and Nason (2011) develop a bootstrap method to deal with the nuisance parameter problem and obtain critical values from the asymptotic distribution. Their bootstrap, closely related to those of White (2000) and Hansen (2005), uses a block bootstrap algorithm to obtain artificial samples of loss differentials from the empirical sample of loss differentials.

5.4 Application

To illustrate the testing of equal (population-level) forecast accuracy in multiple nested models, we use an expanded set of the inflation models used in section 3, and the recursive estimation scheme. We take as a benchmark the autoregressive model of equation (14). We consider seven alternative models that generalize equation (15) to include various combinations of GDP growth, the GDP gap, capacity utilization in manufacturing, and a measure of food and energy inflation, defined as the difference between overall and ex food and energy inflation in the price index for personal consumption expenditures (using four-quarter rates of inflation). The variable combinations are listed in the first column of Table 6. The second column of the table gives the RMSE of each model relative to the benchmark RMSE. As the RMSE ratios indicate, most of the models are less accurate than the AR benchmark. The only exception is the model that uses GDP growth to forecast inflation, for which the RMSE ratio is 0.980 at the one-quarter horizon and 0.920 at the four-quarter horizon. Accordingly, this model is best at both forecast horizons.

To test the predictive ability at the population level, one possible approach would be to use the bootstrap methods and tests of White (2000) or Hansen (2005). However, because the alternative models nest the benchmark model, such an approach would not be valid. In practice, if applied to nested models anyway, White's (2000) reality check or Hansen's (2005) test of superior predictive ability (SPA) will tend to yield few rejections (Monte Carlo evidence in Clark and McCracken (2011b) shows these tests to be significantly undersized). In fact, in this application, the biggest MSE-t test statistics (given in the last column of the table) are just 0.898 at the 1-quarter horizon and 0.564 at the 4-quarter horizon; the corresponding pairwise normal-based p-values are 0.185 and 0.287, respectively (Table 3).

As a result, it is very unlikely that White's (2000) reality check or Hansen's (2005) SPA test would yield a rejection.

In this application, the better, asymptotically valid approach to testing equal accuracy in population would consist of using MSE-F and MSE-t statistics (provided in the last two columns of Table 2) and p-values obtained with the fixed regressor bootstrap of Clark and McCracken (2011b). Accordingly, the first row of each panel in Table 6 (which includes one panel for the one-quarter forecast horizon and another for the four-quarter horizon) provides the MSE-F and MSE-t test statistics for the best of the alternative models, along with the Clark-McCracken reality check p-values. The remaining rows provide the pairwise p-values for each test against the benchmark, obtained with the fixed regressor bootstrap under the null of equal predictive ability at the population level.

On a pairwise basis, the null of equal accuracy at the population level cannot be rejected for most models, with the single exception of the Phillips Curve that includes GDP growth.[24] For this model, the more powerful MSE-F test rejects the null at both forecast horizons (at a confidence level of 5 percent), while the MSE-t test rejects the null (at a confidence level of 10 percent) at the one-quarter horizon but not the four-quarter horizon. As expected, for this best model, the reality check p-values are somewhat higher than the pairwise p-values. Still, using a significance level of 10 percent, the reality check version of the MSE-F test rejects the null of equal accuracy at both forecast horizons. However, the reality check version of the MSE-t test does not reject the null at either horizon. Overall, based on the power differences of the tests, it seems reasonable to conclude that, at a population level, the Phillips Curve with GDP growth forecasts significantly better than the AR model. This is true on a pairwise basis — a basis also considered in section 3, with the same conclusion — and on a multiple-model basis.

6 Evaluation of Real-Time Forecasts

Throughout the literature on forecast evaluation, one issue that is almost always overlooked is the real-time nature of the data being used. For example, in section 2 we laid out a framework for forecasting for which, at each forecast origin $t = R, \ldots, T - \tau$, we observe a sequence of observables $\{y_s, x_s'\}_{s=1}^{t}$ that includes a scalar random variable y_t to be predicted,

[24] The pairwise bootstrap p-values in Table 6 differ slightly from those given in the "FRBS, population EPA" results of Table 4 due to differences in random numbers associated with generating the results separately.

as well as a $(k \times 1)$ vector of predictors x_t. In particular, note that the notation being used implies that the difference between the information sets at time t and time $t+1$ consists exclusively of the pair $\{y_{t+1}, x'_{t+1}\}$. This framework for forecasting makes perfect sense in the cases when both y and x consist of unrevised financial variables like interest and exchange rates. Hence for many financial applications, including Goyal and Welch (2008) or Chen, Rogoff, and Rossi (2010), this framework is perfectly reasonable.

But once we start looking into the predictive content of macroeconomic variables, the use of this framework becomes tenuous due to the fact that as we move across forecast origins, the historical values of many macroeconomic series (including GDP, employment, and to a somewhat lesser degree inflation) are revised. In order to capture this feature, consider instead a framework for forecasting for which, at each forecast origin $t = R, \ldots, T - \tau$, we observe a sequence of observables $\{y_s(t), x'_s(t)\}_{s=1}^{t}$ that includes a scalar random variable $y_s(t)$ to be predicted, as well as a $(k \times 1)$ vector of predictors $x_s(t)$. As was the case above, the subscript continues to denote the historical date associated with the value of the variable but now we have the parenthetical (t). This additional notation is intended to make clear that as statistical agencies gather more data across time, and sometimes even change the definitions of variables, the historical value of a particular variable can change. In other words, the difference between the information sets at time t and time $t+1$ consists not only of the pair $\{y_{t+1}(t+1), x'_{t+1}(t+1)\}$ but potentially the entire sequence of past observables.

There are several ways around this issue when it comes to out-of-sample forecast evaluation. The easiest and most common approach is to ignore the real-time issue. For example, Stock and Watson (2003) conduct pseudo out-of-sample forecasting exercises designed to look at the predictive content of asset prices for a variety of macroeconomic series. In that exercise they use 2000 to 2002-vintage macroeconomic data.[25] In their exercise, they — like most other researchers in the forecasting literature (including ourselves, in some other papers) — completely ignore the possibility that the data has been revised across time. By taking that approach they do not truly address the question of whether asset prices have predictive content for macroeconomic series so much as they address a related question: Would asset prices have had predictive content for macroeconomic variables if the present vintage of data had been available historically at each forecast origin $t = R, \ldots, T - \tau$? To be fair, Stock and Watson were well aware of this issue. They provide a rationale for their

[25] For example, the GDP-related files in the dataset Mark Watson has kindly made publicly available have date stamps of May 20, 2000. The files for other variables have date stamps ranging up to late 2002.

choice in footnote 3 of the corresponding paper.

A second, subtle approach is advocated by Koenig, Dolmas and Piger (2003). They suggest using the various vintages of data as they would have been observed in real time to construct forecasts. In the notation above they advocate conducting the pseudo out-of-sample forecast exercise only using the values of the series observed at the time that the forecast was constructed. In this framework the only relevant data at each forecast origin $t = R, \ldots, T - \tau$ consist of the observables$\{y_s(s), x_s'(s)\}_{s=1}^{t}$. Were we to take this approach, the additional parentheticals (s) become vacuous and we revert to the framework discussed throughout this chapter. Clements and Galvao (2010) apply the approach of Koenig, Dolmas and Piger (2003) to forecasting GDP growth and inflation with AR models.

A final, and much more difficult approach is not to ignore the revision process across vintages of the macroeconomic series and to deal with the vintages of data in the way they are most commonly used. In this approach the pseudo out-of-sample forecasting exercise explicitly takes into account the fact that the values of the reported y and x variables may vary across time. As shown in Clark and McCracken (2009) this may very well lead to differences in the statistical behavior of out-of-sample tests of predictive ability. This arises because by their nature, out-of-sample tests are particularly susceptible to changes in the correlation structure of the data as the revision process unfolds. This susceptibility has three sources: (i) while parameter estimates are typically functions of only a small number of observations that remain subject to revision, out-of-sample statistics are functions of a sequence of parameter estimates (one for each forecast origin), (ii) the predictand used to generate the forecast and (iii) the dependent variable used to construct the forecast error may be subject to revision and hence a sequence of revisions contribute to the test statistic. If data subject to revision possess a different mean and covariance structure than final revised data (as Aruoba 2008 finds), tests of predictive ability using real-time data may have a different asymptotic distribution than tests constructed using data that is never revised.

The issue is of increasing importance for a couple of reasons. First, as shown in Diebold and Rudebusch (1991), Amato and Swanson (2001), Christoffersen, Ghysels, and Swanson (2002), and Orphanides and van Norden (2005), the predictability of various models is often very different when using real-time vintages of data instead of using the most recent final-vintage data. And second, real-time vintages of macroeconomic data are becoming

increasingly available not only for the U.S. but also for a range of other economies.[26] This has made it much easier for researchers who are interested in forecasting to conduct their pseudo out-of-sample forecasting exercises in a fashion that is significantly closer to the real-world in which policy makers have to construct forecasts and make decisions based upon them.

Of course, one might wonder why the data used in forecast evaluation should be real-time, and why forecasts aren't constructed taking revisions into account. Stark and Croushore (2003) argue forecasts should be evaluated with real-time data because practical forecasting — especially from the standpoint of a policy maker who has to make decisions based upon said forecasts — is an inherently real-time exercise. Reflecting such views, the number of studies using real-time data in forecast evaluation is now quite large (see, e.g., the work surveyed in Croushore (2006) and the list Dean Croushore kindly maintains at https://facultystaff.richmond.edu/ dcrousho/data.htm). As to the construction of forecasts, Croushore (2006) notes that, in the presence of data revisions, the optimal approach will often involve jointly modeling the final data and revision process, and forecasting from the resulting model (e.g., Howrey 1978, Kishor and Koenig (2011)).

More commonly, though, forecasts are generated at a moment in time using the most recent vintage of data. Accordingly, Clark and McCracken (2009) focus on such an approach, and provide results covering the most common practices: generating forecasts with real-time data and evaluating the forecasts with either preliminary or final data. To accomplish this they make a simplifying assumption about the revision process. In particular they assume that macroeconomic series are revised for a finite number of periods r (which they refer to as the "vintage horizon"), after which the series are not revised.[27] In this framework, at each forecast origin we continue to observe a sequence of observables$\{y_s(t), x_s'(t)\}_{s=1}^t$ that are subject to revision across forecast origins with the caveat that for all $t \geq s+r$, $y_s(t) = y_s$ and $x_s(t) = x_s$: The parenthetical is dropped when the revision process is completed.

As an example, consider the case in which the predictive content of two linear models $y_{s+\tau}(t) = x_{1,s}'(t)\beta_1^* + u_{1,s+\tau}(t)$ (model 1) and $y_{s+\tau}(t) = x_{2,s}'(t)\beta_2^* + u_{2,s+\tau}(t)$ (model 2) are being compared. For each forecast origin t the variable to be predicted is $y_{t+\tau}(t')$,

[26] Data for the U.S. are readily accessible at the Federal Reserve Banks of Philadelphia (http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/) and St. Louis (http://research.stlouisfed.org/tips/alfred/). See Dean Croushore's website for a more complete list of U.S. and international data sources: https://facultystaff.richmond.edu/ dcrousho/data.htm.

[27] Annual and benchmark revisions are ignored.

where $t' \geq t + \tau$ denotes the vintage used to evaluate the forecasts. In the context of one quarter-ahead forecasts of GDP growth y_{t+1}, this vintage may be the initial release at the end of the first month following the end of the present quarter ($y_{t+1}(t+1+1 \text{ month})$), may be the first revised value at the end of the second month following the end of the quarter ($y_{t+1}(t+1+2 \text{ months})$), or the final release at the end of the third month following the end of the present quarter ($y_{t+1}(t+1+3 \text{ months})$).

For fixed values of the vintage horizon r and the vintage t' used to evaluate the forecasts, Clark and McCracken (2009) revisit the asymptotic theory for population-level tests of equal forecast accuracy between these two OLS-estimated models when they are non-nested or nested models. They find that whether or not the standard asymptotics discussed in sections 3.1.1 and 3.1.2 continue to apply depends critically upon the properties of the data revisions.

6.1 Non-nested comparisons

As we did in section 3.1.1, consider a test of equal MSE based upon the sequence of loss differentials $\hat{d}_{t+\tau}(t') = \hat{u}^2_{1,t+\tau}(t') - \hat{u}^2_{2,t+\tau}(t')$. In a framework with data revisions, Clark and McCracken (2009) show that West's (1996) result of asymptotic normality and asymptotically-irrelevant estimation risk (making $\Omega = S_{dd}$) can break down. In particular they show that if the data revisions are predictable, the statistic

$$\text{MSE-}t = (P - \tau + 1)^{1/2} \frac{\bar{d}}{\sqrt{\hat{\Omega}}}. \tag{25}$$

is asympotically standard normal where, with a proper redefinition of terms, Ω takes the form presented in equation (2) of section 3.1.1. Specifically

$$\Omega = S_{dd} + 2\lambda_{fh}(FBS'_{fh}) + \lambda_{hh}FBS_{hh}B'F', \tag{26}$$

with $F = (-2Eu_{1,t+\tau}(t')x'_{1,t}(t), 2Eu_{2,t+\tau}(t')x'_{2,t}(t))$, B a block diagonal matrix with block diagonal elements B_1 and B_2, S_{dd} the long-run variance of $d_{t+\tau}(t')$, S_{hh} the long-run variance of $h_{t+\tau}$, and S_{dh} the long-run covariance of $h_{t+\tau}$ and $d_{t+\tau}$.

Since the asymptotic variance Ω has the same form as that in West (1996), some of the special cases in which one can ignore parameter estimation error remain the same. For example, if the number of forecasts $P - \tau + 1$ is small relative to the number of in-sample observations from the initial forecast origin R, such that $\pi = 0$, then λ_{fh} and λ_{hh} are zero and hence the latter covariance terms are zero.

Another special case arises when F equals zero. In this case the latter covariance terms are zero and hence parameter estimation error can be ignored. To see when this will or will not arise it is useful to write out the population forecast errors explicitly. That is, consider the moment condition $E(y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x'_{i,t}(t)$. Moreover, note that β_i^* is defined as the probability limit of the regression parameter estimate in the regression $y_{s+\tau} = x'_{i,s}\beta_i^* + u_{i,s+\tau}$. Hence F equals zero if $Ex_{i,t}(t)y_{t+\tau}(t') = (Ex_{i,t}(t)x'_{i,t}(t))(Ex_{i,t}x'_{i,t})^{-1}(Ex_{i,t}y_{t+\tau})$ for each $i = 1, 2$. Some specific instances that result in $F = 0$ are listed below.

1. x and y are unrevised.
2. x is unrevised and the revisions to y are uncorrelated with x.
3. x is unrevised and final revised vintage y is used for evaluation.
4. x is unrevised and the "vintages" of y's are redefined so that the data release used for estimation is also used for evaluation (as suggested by Koenig, Dolmas and Piger (2003)).

In general, though, neither of these special cases — that $\pi = 0$ or $F = 0$ — need hold. In the former case, West and McCracken (1998) emphasize that in finite samples the ratio $P/R = \widehat{\pi}$ may be small but that need not guarantee that parameter estimation error is negligible since it may be the case that $FBS_{dh} + FBS_{hh}BF'$ remains large. For the latter case, in the presence of predictable data revisions it is typically not the case that $F = 0$. To conduct inference then requires constructing a consistent estimate of the asymptotic variance Ω.

6.2 Nested comparisons

In section 3.1.2, we showed that tests of equal population-level predictability between nested models have asymptotic distributions that are typically non-standard — that is, not asymptotically standard normal or χ^2. However, these results required the absence of data revisions. In the presence of *predictable data revisions*, the asymptotics for these tests change dramatically — much more so than in the non-nested case.[28] The key issue in the analytics is that when there are data revisions, the residuals $y_{s+\tau} - x'_{i,s}\beta_i^*$, $s = 1, \ldots, t - \tau$, and the forecast errors $y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*$, $t = R, \ldots, T - \tau$, need not have the same covariance structure.

Keeping track of this distinction, Clark and McCracken (2009) show that for nested

[28] Mankiw, Runkle, and Shapiro (1984) refer to predictable revisions as "noise" and unpredictable revisions as "news."

model comparisons the statistic

$$\text{MSE-}t = (P - \tau + 1)^{1/2} \frac{\bar{d}}{\sqrt{\hat{\Omega}}} \tag{27}$$

is asympotically standard normal, where Ω takes the form

$$\Omega = \lambda_{hh} F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F', \tag{28}$$

with $F = 2Eu_{2,t+\tau}(t')x'_{2,t}(t)$ and B_1, B_2, S_{hh} as defined in section 3.1.2.

The result makes clear that in the presence of predictable revisions, a t-test for equal predictive ability can be constructed that is asymptotically standard normal under the null hypothesis — even when the models are nested. This is in sharp contrast to the results in Clark and McCracken (2005a) and McCracken (2007), in which the tests generally have non-standard limiting distributions. This finding has a number of important implications, listed below.

1. The statistic MSE-$t = (P-\tau+1)^{1/2}\bar{d}/\sqrt{\hat{S}_{dd}}$ diverges with probability 1 under the null hypothesis. This occurs because (i) $(P-\tau+1)^{1/2}\bar{d}$ is asymptotically normal and (ii) $\hat{S}_{dd}$ is a consistent estimate of S_{dd}, which is zero when the models are nested. A similar argument implies the MSE-F statistic also diverges with probability 1 under the null hypothesis.

2. Out-of-sample inference for nested comparisons can be conducted without the strong auxiliary assumptions made in Clark and McCracken (2005a) and McCracken (2007) regarding the correct specification of the models. Optimal forecasts from properly specified models will generally follow an MA($\tau - 1$) process, which we typically required in our prior work. In the presence of predictable revisions, the serial correlation in τ-step forecast errors can take a more general form.

3. Perhaps most importantly, asymptotically valid inference can be conducted without the bootstrap or non-standard tables. So long as an asymptotically valid estimate of Ω is available, standard normal tables can be used to conduct inference. Consistent methods for estimating the appropriate standard errors are described in section 3.1.1.

Regardless, it is possible that the asymptotic distribution of the MSE-t test can differ from that given in equations (27) and (28). The leading case occurs when the revisions are *unpredictable* rather than predictable, so that $F = 2Eu_{2,t+\tau}(t')x'_{2,t}(t) = 0$. Another occurs when model 1 is a random walk and model 2 includes variables subject to predictable revisions. But even with predictable revisions that make F non-zero, asymptotic normality fails to hold when $F(-JB_1J' + B_2)$ (and hence Ω) equals zero. In both cases Clark and

McCracken (2009) establish that the MSE-t statistic (from (27)) is bounded in probability under the null. However, in each instance the asymptotic distributions are non-standard in much the same way as the results in Clark and McCracken (2005a). Moreover, conducting inference using these distributions is complicated by the presence of unknown nuisance parameters. A complete characterization of these distributions has yet to be delineated.

6.3 Application

To illustrate the testing of equal (population-level) forecast accuracy in real-time forecasts, we use the same set of inflation models as in section 3, but with real-time data on GDP and the GDP price index, obtained from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists (RTDSM). The full forecast evaluation period runs from 1985:Q1 through 2008:Q4. For each forecast origin t in 1985:Q1 through 2008:Q4, we use data vintage t to estimate the output gap, (recursively) estimate the forecast models, and then construct forecasts for periods t and beyond. We treat the time series of trend inflation as unrevised throughout the analysis.[29] The starting point of the model estimation sample is always 1962:2. Following the example in Clark and McCracken (2009), in evaluating forecast accuracy, we consider several possible definitions (vintages) of actual inflation. One estimate is the second one available in the RTDSM, published two quarters after the end of the forecast observation date. We also consider estimates of inflation published with delays of five and 13 quarters.

The top panel of Table 7 presents results for the non-nested comparison of forecasts from the models with the output gap (model 1) and GDP growth (model 2). In terms of MSEs, in contrast to the previous results on forecast accuracy in current vintage data, for some horizons (one-quarter) and definitions of actual inflation, the model with the output gap yields forecasts more accurate than does the model with GDP growth. However, there is little evidence of statistical significance in any of the (non-nested) forecast accuracy differences. This is true when the test statistics are based on the conventional variance $\widehat{S}_{dd}$ and when the test statistics are based on the adjusted, larger variance $\widehat{\Omega}$ (which takes account of the potential for predictability in the data revisions); in this application, as in Clark and McCracken (2009), correcting the standard error for the predictability of data revisions doesn't have much impact on the test result. Overall, in this non-nested model

[29] This is appropriate for the survey-based portion of the trend series. To the extent that definitional changes in actual inflation across vintages affect the average inflation rate, our inclusion of an intercept in all of the forecasting models suffices to capture these differences in inflation levels.

comparison, using testing methods robust to data revisions does not change the current-vintage application result of section 3, in which the null of equal accuracy (at the population level) cannot be rejected.

The second and third panels of Table 7 provide results for the nested model comparison of forecasts from Phillips Curve models versus the AR benchmark. The MSEs indicate that, in these real-time forecasts, the Phillips Curve is almost always more accurate than the benchmark. When we abstract from the potential impact of predictable data revisions on test behavior, and compare MSE-F and MSE-$t(S_{dd})$ to asymptotic critical values simulated as in Clark and McCracken (2005), we almost always reject the null of equal accuracy, for each Phillips Curve specification against the AR benchmark.[30] The one exception is for the Phillips Curve using the GDP gap at the four-quarter ahead horizon and the estimate of GDP published with a 13 period delay to measure actual inflation. As might be expected based on the results of Clark and McCracken (2009), taking account of data revisions by using the variance $\widehat{\Omega}$ (rather than the conventional variance $\widehat{S}_{dd}$) in the MSE-t test always increases the absolute value of the t-statistic. However, there are only a handful of cases in which the adjusted t–statistic compared against Clark-McCracken critical values is significant when the unadjusted t–statistic (compared against standard normal critical values) is not. In this application, for a Phillips curve with GDP growth, the evidence of predictive content at the population level is about the same in real time as in final vintage data (section 3), while for a Phillips curve with the GDP gap, the evidence of predictive content is somewhat stronger in the real time data than in final vintage data.

7 Small-Sample Properties of Tests of Equal Predictive Ability

In this section we review the small-sample properties of the testing methods reviewed in sections 3-5, first summarizing existing findings and then presenting a new Monte Carlo comparison of alternative HAC estimators in nested model forecast evaluation. We also describe a theory-based approach to including a size correction in some test statistics.

Most recent assessments of the small-sample behavior of tests of predictive ability applied to pairs of forecasts have focused on forecasts from nested models. Accordingly, our survey

[30] Throughout these real-time examples, in computing the MSE-t tests, we use the Newey and West (1987) estimator of the necessary long-run variances, with a bandwidth of 2 at the one-quarter forecast horizon and 8 at the four-quarter horizon

of evidence on small-sample properties focuses on nested model comparisons. For evidence on the properties of tests applied to forecasts from non-nested models or forecasts that don't involve model estimation, see such studies as Clark (1999), Diebold and Mariano (1995), McCracken (2000), West (1996), and Busetti, Marcucci, and Veronese (2009).

For tests of equal predictive ability at the population level, Monte Carlo results in Clark and McCracken (2001, 2005a), Clark and West (2006, 2007), and McCracken (2007) show that critical values obtained from Monte Carlo simulations of the asymptotic distributions generally yield good size and power properties for 1-step ahead forecasts, but can yield rejection rates greater than nominal size for multi-step forecasts. Similarly, results in Clark and West (2006, 2007) indicate that comparing the ENC-t or Clark-West test against standard normal critical values can work reasonably well but exhibit size distortions as the forecast horizon increases (note that, for null models that take a random walk form, these distortions can be avoided by using the Hodrick (1992) estimator of the standard deviation that enters the test statistic). Later in this section we examine whether the size performance of the ENC-t test based on normal critical values can be improved by using an alternative HAC estimator of the standard error in the denominator of the test statistic.

A number of Monte Carlo studies have shown that some bootstrap approaches can yield good size and power properties for tests of equal predictive ability at the population level. Clark and McCracken (2001, 2005a) and Clark and West (2006, 2007) find that the restricted VAR bootstrap described in section 3.1.2 works well in a range of settings. Experiments in Clark and McCracken (2011a, 2011b) and section 7.1 below show that the fixed regressor bootstrap under the null of equal predictive ability at the population level (also referred to as a no-predictability fixed regressor bootstrap) works equally well. Both of these bootstrap approaches offer the advantage that they yield accurately sized tests even at long forecast horizons.

For tests of equal predictive ability in a finite sample, Giacomini and White (2006) present Monte Carlo evidence that, for 1-step ahead forecasts generated under a rolling estimation scheme, comparing a t-test for equal MSE against standard normal critical values has reasonable size and power properties. However, their results are based on two-sided tests. If a researcher or practitioner prefers to take the smaller forecasting model as the null to be rejected only if it is less accurate than the larger model (as opposed to also rejecting the larger model in favor of the smaller), he or she would consider a one-sided

test. Examining this case, Clark and McCracken (2011a, 2011c) find that comparing t-tests of equal MSE against standard normal critical values (under a null of equal accuracy in the finite sample) tends to yield modestly under-sized tests, especially at shorter forecast horizons. The under-sizing is actually a bit worse with forecasts generated under a rolling estimation scheme than under a recursive scheme, even though the former is justified by the results of Giacomini and White and the latter is not. One other puzzle highlighted in Clark and McCracken's (2011a, 2011c) Monte Carlo analysis across a wide range of sample sizes is that, when the MSE-t test is compared against standard normal critical values, the rejection rate falls as P/R rises. This pattern runs contrary to the asymptotic results of Giacomini and White (2006), which imply that the test should be more accurate when P is large. It is possible, of course, that the asymptotics kick in very slowly.

Clark and McCracken (2011a) find that comparing tests of equal MSE against critical values generated from a pairwise simplification of White's (2000) non-parametric bootstrap yields results very similar to those obtained for standard normal critical values — consistent, although sometimes just modest, undersizing. Corradi and Swanson (2007) also generally find the non-parametric bootstrap to be under-sized when applied to 1-step ahead forecasts from nested models. White's bootstrap offers the advantage of simplicity, as it only involves re-sampling forecast errors. While White showed the bootstrap to be asymptotically valid for non-nested models, the bootstrap may be valid under the asymptotics of Giacomini and White (2006), for forecasts generated from an estimation sample of a fixed size (rolling window estimation scheme).

For a range of DGPs and settings, the Monte Carlo evidence in Clark and McCracken (2011a, 2011c) shows that, for testing equal forecast accuracy in the finite sample, the fixed regressor bootstrap detailed in section 3.2.2 works well. When the null of equal accuracy in the finite sample is true, the testing procedures yield approximately correctly sized tests. When an alternative model is, in truth, more accurate than the null, the testing procedures have reasonable power. However, using this bootstrap at longer forecast horizons tends to result in some over-sizing, stemming from imprecision in the HAC estimate of the variance matrix V used to determine the parameterization of the bootstrap DGP. In the next section, we consider whether alternative HAC estimators improve the reliability of the bootstrap at longer forecast horizons.

As to small sample properties in tests of multiple forecasts, Hubrich and West (2010)

show their proposed maximum Clark-West test to be slightly undersized and the χ^2 test based on the Clark-West numerators to be slightly oversized, when applied to 1-step ahead forecasts from 3 or 5 models. The maximum test has better power than the χ^2 test. For comparison, Hubrich and West also provide results based on White's (2000) non-parametric reality check bootstrap, which is asymptotically valid for non-nested models (under some additional conditions) but not nested models. They find the reality check to be somewhat undersized, or even severely undersized in small samples. For the maximum and χ^2 based on Clark-West-adjusted loss differentials, Granziera, Hubrich, and Moon (2011) obtain similar Monte Carlo results for forecasts from 3 or 4 models. Their proposed likelihood ratio test improves on the finite-sample power of the Hubrich-West χ^2 test, but the power rankings of the likelihood ratio test and maximum Clark-West test vary with the application setting and sample size. Granziera, Hubrich, and Moon (2011) find tests based on the fixed regressor bootstrap of Clark and McCracken (2011b) to be slightly undersized to correctly sized.

Clark and McCracken (2011b) provide Monte Carlo results for experiments with much larger numbers of forecasts (experiments with 17 and 128 models) and both a 1-step and 4-step ahead forecast horizon. They find that tests of equal MSE and forecast encompassing based on the fixed regressor bootstrap have good size properties (i.e., have empirical size close to nominal size) in a range of settings. But they also show that, in applications with high persistence in predictors and high correlations between innovations to the predictand and the predictors (so that the problems highlighted by Stambaugh (1999) apply), the tests can be modestly oversized. Under general conditions, in most, although not all, cases, the tests of forecast encompassing have slightly lower size than tests of equal MSE. In broad terms, the F-type and t-type tests have comparable size. Considering other testing approaches, Clark and McCracken find that, in experiments with 17 forecasting models, comparing the ENC-t (or Clark-West) test against critical values obtained with the Hubrich and West (2010) approach have reasonable size properties at the 1-step horizon, but not the 4-step horizon, especially in small samples. Multi-step size distortions are smaller in the simulation results of Granziera, Hubrich, and Moon (2011), which involve fewer models. The over-sizing appears to be due to small-sample imprecision of the autocorrelation-consistent estimated variance of the normal random variables, obtained as in Newey and West (1987); perhaps other HAC estimators could reduce the size distortions. Finally, consistent with the evidence in Hubrich and West (2010), Clark and McCracken find that tests of equal MSE

based on critical values obtained from White's (2000) non-parametric bootstrap are generally unreliable — for the null of equal accuracy at the population level — in application to nested models. Rejection rates based on the non-parametric bootstrap are systematically too low in size experiments and lower than rates based on other approaches in power experiments. Corradi and Swanson (2007) report similar results for some other tests of equal predictive ability, applied to pairs of nested models.

7.1 Monte Carlo Comparison of Alternative HAC Estimators, in Pairs of Models

In practice, one unresolved challenge in forecast test inference is achieving accurately sized tests applied at multi-step horizons — a challenge that increases as the forecast horizon grows and the size of the forecast sample declines. The root of the challenge is precise estimation of the HAC variance that enters the test statistic. For example, in Clark and McCracken's (2005a) Monte Carlo assessment of the properties of tests of equal accuracy in population, using asymptotic critical values yields size distortions that increase with the forecast horizon and can be substantial in small samples. Bootstrapping the test statistic can effectively deal with the problem: as documented in sources such as Clark and McCracken (2005a), comparing the same tests against bootstrapped critical values yields accurately sized tests.

However, bootstrap methods are not necessarily a universal solution. One reason noted above, is that, for tests of the null of equal accuracy in the finite sample, Clark and McCracken (2011a) find that the use of a bootstrap is by itself not enough to eliminate size distortions. A second reason is that, to avoid the computational burden of bootstrapping critical values, some researchers may prefer to construct test statistics that can be compared against asymptotic critical values without size distortions. For example, in applications that involve using the test of Clark and West (2006, 2007) to test equal forecast accuracy in population, some might find it helpful to be able to compare some version of the test against the Clark and West-suggested normal critical values, without the problem of sharp size distortions at multi-step horizons.

Some past research suggests that judicious choice of the HAC estimator could improve size performance at longer forecast horizons. Most past work on the finite-sample properties of forecast tests has used the HAC estimator of Newey and West (1987), seemingly the most common HAC estimator in empirical work. However, Clark and West (2006) find that using

the HAC estimator of Hodrick (1992) — which can be applied with a martingale difference null, but not with more general null models — yields much better size properties for their proposed test of equal forecast accuracy. The results of Harvey, Leybourne, and Newbold (1997) also suggest that, in some cases, test size could be improved by making a simple finite-sample adjustment to the test.

Building on this past work, in this section we conduct a systematic Monte Carlo examination of whether alternative HAC estimators can alleviate size distortions that can arise with the estimator of Newey and West (1987). We focus on tests applied to forecasts from nested models, under the null of equal accuracy in population and under the null of equal accuracy in the finite sample. Drawing on the setup of Clark and McCracken (2011a), we use simulations of bivariate and multivariate DGPs based on common macroeconomic applications. In these simulations, the benchmark forecasting model is a univariate model of the predictand y; the alternative models add lags of various other variables of interest.

With data simulated from these processes, we form three basic test statistics using a range of HAC estimators and compare them to alternative sources of critical values. The first subsection details the data-generating processes. The next subsection describes the alternative HAC estimators. The following subsection lists the sources of critical values. Remaining subsections present the results. We focus our presentation on recursive forecasts, and we report empirical rejection rates using a nominal size of 10%.

7.1.1 Monte Carlo design

For all DGPs, we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP. We consider forecast horizons of four and eight steps. Note that, in this Monte Carlo analysis, to facilitate comparisons across forecast horizons, for a forecast horizon of τ, we report results for samples of $\tilde{P} = P + \tau - 1$ forecasts, so that the number of forecasts is the same for each τ. With quarterly data in mind, we also consider a range of sample sizes $(R, \tilde{P})$, reflecting those commonly available in practice: 40,80; 40,120; 80,20; 80,40; 80,80; 80,120; 120,40; and 120,80.

The two DGPs we consider are based on empirical relationships among U.S. inflation and a range of predictors, estimated with 1968-2008 data. In all cases, our reported results are based on 5000 Monte Carlo draws and, with bootstrap methods, 499 bootstrap replications.

DGP 1 is based on the empirical relationship between the change in core PCE inflation (y_t) and the Chicago Fed's index of the business cycle ($x_{1,t}$, the CFNAI), where the change

in inflation is the change in the four-quarter rate of inflation:[31]

$$\begin{aligned} y_{t+\tau} &= b_{11}x_{1,t} + v_{t+\tau} \\ v_{t+\tau} &= \varepsilon_{t+\tau} + \sum_{i=1}^{\tau-1} \theta_i \varepsilon_{t+\tau-i} \\ (\theta_1, \ldots, \theta_{\tau-1}) &= (0.95, 0.9, 0.8) \text{ for } \tau = 4 \\ (\theta_1, \ldots, \theta_{\tau-1}) &= (0.90, 0.95, 0.95, 0.65, 0.6, 0.5, 0.4) \text{ for } \tau = 8 \\ x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\ \text{var}\begin{pmatrix} \varepsilon_{t+\tau} \\ v_{1,t+\tau} \end{pmatrix} &= \begin{pmatrix} 0.2 & \\ 0.0 & 0.3 \end{pmatrix} \text{ for } \tau = 4 \\ \text{var}\begin{pmatrix} \varepsilon_{t+\tau} \\ v_{1,t+\tau} \end{pmatrix} &= \begin{pmatrix} 0.5 & \\ 0.0 & 0.3 \end{pmatrix} \text{ for } \tau = 8. \end{aligned} \tag{29}$$

In the DGP 1 experiments, the forecasting models are:

$$\text{null: } y_{t+\tau} = \beta_0 + u_{1,t+\tau} \tag{30}$$

$$\text{alternative: } y_{t+\tau} = \beta_0 + \beta_1 x_{1,t} + u_{2,t+\tau}. \tag{31}$$

We consider experiments with different settings of b_{11}, the coefficient on $x_{1,t}$, chosen to reflect particular null hypotheses. First, the coefficient is set to 0, to assess tests of the null of equal forecast accuracy in population. Second, the coefficient is set to a value that makes the models equally accurate (in expectation) on average over the forecast sample. To determine the coefficient value, we begin with an (empirically-based) coefficient of $b_{11} = 0.4$ for $\tau = 4$ and $b_{11} = 1.0$ for $\tau = 8$. For each $R, \tilde{P}$ combination, we use the asymptotic theory of Clark and McCracken (2011a) to determine a preliminary re-scaling of the coefficient to yield equal accuracy. For each $R, \tilde{P}$ combination, we then conduct three sets of Monte Carlo experiments (with a large number of draws), searching across grids of the re-scaling of the coefficient to select a scaling that minimizes the average (across Monte Carlo draws) difference in MSEs from the competing forecasting models.[32]

[31] Specifically, in the empirical estimates underlying the DGP settings, we defined $y_{t+\tau} = 100\ln(p_{t+\tau}/p_{t+\tau-4}) - 100\ln(p_t/p_{t-4})$, where p denotes the core PCE price index.

[32] Specifically, we first consider 11 different experiments, each using 20,000 draws and a modestly different set of coefficient values obtained by scaling the baseline values, using a grid of scaling factors. We then pick the coefficient scaling that yields the lowest (in absolute value) average (across draws) difference in MSEs. We then repeat the 11-experiment exercise. Finally, we consider a third set of 21 experiments, with a more refined grid of coefficient scaling values and 200,000 draws. The coefficient scaling value that yields the smallest (absolute) difference in MSEs in this third set of experiments is then used to set the coefficients in the DGP simulated for the purpose of evaluating test properties.

DGP 2 extends DGP 1 to include more predictands for y:

$$\begin{aligned}
y_{t+\tau} &= b_{11}x_{1,t} + b_{21}x_{2,t} + b_{31}x_{3,t} + v_{t+\tau} \\
v_{t+\tau} &= \varepsilon_{t+\tau} + \sum_{i=1}^{\tau-1} \theta_i \varepsilon_{t+\tau-i} \\
(\theta_1, \ldots, \theta_{\tau-1}) &= (0.95, 0.9, 0.8) \text{ for } \tau = 4 \\
(\theta_1, \ldots, \theta_{\tau-1}) &= (0.90, 0.95, 0.95, 0.65, 0.6, 0.5, 0.4) \text{ for } \tau = 8 \\
x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\
x_{2,t+1} &= 0.8x_{2,t} + v_{2,t+1} \\
x_{3,t+1} &= 0.8x_{3,t} + v_{3,t+1} \\
\text{var}\begin{pmatrix} \varepsilon_{t+\tau} \\ v_{1,t+\tau} \\ v_{2,t+\tau} \\ v_{3,t+\tau} \end{pmatrix} &= \begin{pmatrix} 0.2 & & & \\ -0.01 & 0.3 & & \\ 0.03 & 0.03 & 2.2 & \\ -0.2 & 0.02 & 0.8 & 9.0 \end{pmatrix} \text{ for } \tau = 4 \\
\text{var}\begin{pmatrix} \varepsilon_{t+\tau} \\ v_{1,t+\tau} \\ v_{2,t+\tau} \\ v_{3,t+\tau} \end{pmatrix} &= \begin{pmatrix} 0.5 & & & \\ 0.05 & 0.3 & & \\ -0.08 & 0.03 & 2.2 & \\ 0.3 & 0.02 & 0.8 & 9.0 \end{pmatrix} \text{ for } \tau = 8.
\end{aligned} \tag{32}$$

In the DGP 2 experiments, the forecasting models are:

$$\text{null: } y_{t+\tau} = \beta_0 + u_{1,t+\tau} \tag{33}$$

$$\text{alternative: } y_{t+\tau} = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_{2,t+\tau}. \tag{34}$$

Again, we consider experiments with different settings of the b_{ij} coefficients, to reflect particular null hypotheses. First, the coefficients are set to 0, to assess tests of the null of equal forecast accuracy in population. Second, the coefficients are set to values that make the competing forecasting models equally accurate (in expectation) on average over the forecast sample. To determine the coefficient vector value, we begin with (empirically-based) coefficients of $b_{11} = 0.4$, $b_{21} = 0.2$, $b_{31} = 0.05$ for $\tau = 4$ and $b_{11} = 1.0$, $b_{21} = 0.2$, $b_{31} = 0.05$ for $\tau = 8$. As described above, for each $R, \tilde{P}$ combination, we use the asymptotic theory of Clark and McCracken (2011a) to determine a preliminary re-scaling of the coefficient vector to yield equal accuracy, and then we conduct three sets of Monte Carlo grid searches to refine the re-scaling that yields (on average) equal forecast accuracy.

7.1.2 Inference approaches

For MSE-F and MSE-t tests of equal MSE and the adjusted t-test of equal MSE developed in Clark and West (2006, 2007), denoted here as CW-t, we consider various HAC estimators

under three different approaches to inference — that is, three different sources of critical values. In all cases, because the competing forecasting models are nested, we only consider one-sided tests, with an alternative hypothesis that the larger forecasting model is more accurate than the smaller.

First, we compare the MSE-t and CW-t tests against standard normal critical values. Under the finite (and fixed) R, large P asymptotics of Giacomini and White (2006), with a null hypothesis of equal accuracy in the finite sample, the MSE-t test applied to rolling forecasts from nested models is asymptotically standard normal. While their result does not apply under a recursive estimation scheme, Clark and McCracken (2011a) find that the size properties of the test are slightly better with recursive forecasts than rolling forecasts. Clark and West (2007) find that, under the null hypothesis of equal accuracy in population, the distribution of the CW-t test (equivalent to the ENC-t test for forecast encompassing considered in such studies as Clark and McCracken (2001, 2005a)) is approximately standard normal (in a range of settings, not necessarily all).

Second, under the null hypothesis of equal accuracy in population, we compare the MSE-F, MSE-t, and CW-t tests against critical values obtained from the no-predictability fixed regressor bootstrap (henceforth, no-predictability FRBS) of Clark and McCracken (2011b). As detailed in section 3.1.4, this bootstrap imposes the null of equal population-level accuracy by restricting β_w to equal 0.

Finally, under the null of equal forecast accuracy in the finite sample, we compare the MSE-F and MSE-t tests against critical values from the fixed regressor bootstrap (henceforth, FRBS) of Clark and McCracken (2011a). As detailed in section 3.2.2, under this procedure, we re-estimate the alternative forecasting model subject to the constraint that implies the null and alternative model forecasts to be equally accurate and generate artificial data, forecasts, and test statistics from this DGP.

7.1.3 HAC estimators

Table 8 lists the alternative HAC estimators we consider with various combinations of the test statistics and sources of critical values.

Following most work in the literature, including our own past Monte Carlo assessments of the small-sample properties of forecast tests, we take the estimator of Newey and West (1987) as the baseline, estimating the variance with 1.5τ lags. While much empirical work fixes the lag length (i.e., the bandwidth), the consistency of the estimator rests on the

Table 8. Alternative HAC Estimators Considered

Estimator	*Source*	*Lags*
NW	Newey and West (1987)	1.5 τ
Rectangular	Hansen (1982)	$\tau - 1$
West	West (1997)	$\tau - 1$
QS	Andrews and Monahan (1992)	data-determined
HLN	Harvey, Leybourne, and Newbold (1997)	$\tau - 1$

bandwidth increasing with sample size. The NW estimator rate converges at a rate of T^{α}, where α is less than 1/2, and $\alpha = 1/3$ if the bandwidth parameter is chosen at the optimal rate developed in Andrews (1991).

One alternative, included in Diebold and Mariano's (1995) original development of the MSE-t test, is the rectangular kernel estimator of Hansen (1982), which exploits or presumes one of the implications of optimality of forecasts, which is serial correlation of order $\tau - 1$. While the Newey-West estimator reduces the weight given to covariances as the lag increases, the rectangular estimator assigns a weight of 1 to all lags up through lag $\tau - 1$. Compared to the NW, West (1997), or quadratic spectral (QS) estimators, the rectangular estimator suffers a disadvantage that it need not be positive semi-definite (in our simulations, in the very rare instance in which that occurred, we replaced the rectangular estimator with the NW estimator). However, compared to the NW and QS estimators, the rectangular estimator converges at a faster rate, of $T^{0.5}$. The imposition of parametric restrictions may offer some gains in small-sample precision over the NW and QS estimators.

We also consider the estimator of West (1997), which generalizes one suggested by Hodrick (1992). Our use of the West estimator is motivated by the Clark and West (2006) finding that, under a martingale difference null that permits the application of Hodrick's (1992) estimator, tests based on Hodrick's HAC estimator have superior size properties. The West estimator involves: fitting an MA model to the residual series of the equation of interest; forming a weighted sum of lags of the right hand side variables from the equation of interest, using the MA coefficients as weights; and then computing the HAC variance as the simple contemporaneous variance of the MA residual times the weighted sum of variables. The West estimator has an advantage over the rectangular estimator of being guaranteed to be positive semi-definite and the advantage over the NW and QS estimators that it converges at a rate of $T^{0.5}$. Again, the imposition of parametric restrictions may offer some gains in small-sample precision over the NW and QS estimators.

Our fourth HAC estimator is the pre-whitened quadratic spectral variance developed by

Andrews and Monahan (1992). For the equation of interest, this estimator involves: pre-whitening the products of the residual and right-hand side variables by fitting a VAR(1); determining the optimal bandwidth for the quadratic spectral kernel to be used with the residuals from the VAR(1); computing the HAC variance for the VAR residuals using this kernel and bandwidth; and then using the VAR structure to compute the HAC variance for the original variables (the products of the residual and right-hand side variables). Compared to the NW estimator, the QS estimator has an advantage in convergence rate. For example, if the bandwidth parameter is chosen at the optimal rate, the QS convergence rate is 2/5, compared to 1/3 for NW. However, the QS estimator is more difficult to compute, particularly with pre-whitening and bandwidth optimization.[33]

Finally, for the MSE-t and CW-t tests compared to standard normal critical values, we consider the adjusted variance developed by Harvey, Leybourne, and Newbold (1997). Their adjustment is a finite-sample one, developed assuming forecasts in which parameter estimator error is irrelevant and the variance is computed with the rectangular estimator included in Diebold and Mariano's (1996) original development of the MSE-t test. The HLN adjustment consists of forming the t-statistic using the rectangular variance estimate and $\tau - 1$ lags and then multiplying the t-test by $\sqrt{\left(\tilde{P} + 1 - 2\tau + \tilde{P}^{-1}\tau\left(\tau - 1\right)\right)/\tilde{P}}$.

In the interest of limiting the volume of results, we limit the combinations of these HAC estimators, test statistics, and inference approaches to the set necessary to determine what must be done to get correctly-sized tests for the relevant null hypothesis.

Under a null of equal accuracy in population, for tests compared against critical values from the no-predictability FRBS, based on prior research the use of the bootstrap is likely to be enough by itself to deliver accurately sized tests. Accordingly, in constructing the MSE-t and CW-t tests for comparison against these critical values, we simply use the NW HAC estimator to compute the denominators of the t-statistics. For the MSE-F test, no HAC estimator enters the computation. With this bootstrap, we don't consider any other HAC estimators.

Under a null of equal accuracy in the finite sample, for tests compared against critical values from the FRBS, the use of the bootstrap isn't enough to deliver accurately sized tests for multi-step forecasts (in small samples), because of imprecision in the HAC variance V

[33] In the interest of brevity, we don't consider the pre-whitened, data-dependent estimator of Newey and West (1994), which uses the Bartlett kernel. In unreported results, Clark and West (2006) found the Andrews and Monahan (1992) estimator to yield slightly to modestly better performance than the Newey and West (1994) estimator.

that plays a role in determining the parameters of the bootstrap DGP. Accordingly, in this case, we consider multiple versions of the bootstrap, each one using a different HAC estimator of V.[34] That is, we generate results for one version of the bootstrap based on the NW estimate of V, another set of results for the bootstrap based on the rectangular estimate of V, and so on.[35] In this case, the computation of the MSE-F and MSE-t tests does not depend on the HAC estimator; for MSE-t, we use the NW variance in the denominator in all cases. Rather, just the bootstrapped data and resulting artificial forecasts, artificial test statistics, and critical values depend on the HAC estimator, through the role of V in determining the DGP.

Finally, for t-tests compared against standard normal critical values, for both of the MSE-t and CW-t statistics, we consider five different versions, each one computed with a different HAC estimate of the standard deviation in the denominator of the t-test. For the occasional Monte Carlo draw in which the rectangular and HLN variances are not positive, we replace the rectangular estimate with the NW estimate of the standard deviation.

7.1.4 Results: Null of equal accuracy in population

We begin with experiments under the null hypothesis of equal forecast accuracy in population, for which Tables 9 and 10 provide Monte Carlo results. Specifically, focusing on the tests and inference approaches that might be expected to yield reasonably-sized tests, Tables 9 and 10 provide results for the MSE-F, MSE-t, and CW-t tests (with the t-statistics computed using the NW estimator) compared against critical values from the no-predictability FRBS and for the CW-t test computed with alternative HAC estimators and compared against standard normal critical values. In light of the common usage of the MSE-t test with normal critical values, we also include results for this test computed with alternative HAC estimators. Under the null of equal accuracy in population, this test should be undersized when compared against standard normal critical values.

The no-predictability FRBS generally yields accurately sized tests. Size peaks at 12% in

[34] To increase the precision of comparisons across HAC estimators, we use the same random numbers to compute results for each different approach to estimating V. Specifically, using the NW estimate of V, we use a random number generator in simulating bootstrap data. We save the underlying random numbers and then use them again when we conduct a bootstrap under the rectangular estimate of V. We proceed to use the same random numbers and conduct bootstraps based on the other estimates of V.

[35] For each alternative approach to estimating V, we follow sources such as Andrews and Monahan (1992) in incorporating a small-sample adjustment. Specifically, we normalize the variance by $T - k$, where k denotes the number of right-hand side variables, rather than T. This small-sample adjustment yields a small, consistent improvement in size.

the experiment with DGP 2, $\tau = 8$, and $R = 120$, $\tilde{P} = 40$. In most other cases, size is quite close to 10%. For example, in the experiment with DGP 2, $\tau = 4$, and $R = 120$, $\tilde{P} = 80$, the sizes of the MSE-F, MSE-t, and CW-t tests are 10.7%, 10.2%, and 9.7%, respectively.

For the CW-t test compared to standard normal critical values, using the NW estimator of the standard deviation in the denominator of the test statistic often, although not always, yields significantly oversized tests — a finding consistent with results in Clark and McCracken (2005a) and Clark and West (2006, 2007). The size distortions increase as the forecast sample shrinks, the forecast horizon rises, and the size of the alternative forecasting model grows. For example, with $R = 120$, $\tilde{P} = 40$, the rejection rate of the NW-based CW-t test is 13.9% with DGP 1 and $\tau = 4$, 18.3% with DGP 1 and $\tau = 8$, 14.9% with DGP 2 and $\tau = 4$, and 20.2% with DGP 2 and $\tau = 8$. With $R = 120$, $\tilde{P} = 80$, the corresponding rejection rates fall to 11.1%, 12.5%, 12.0%, and 15.2%. But in relatively larger forecast samples, shorter forecast horizons, and smaller alternative models, using the NW estimator can yield a reasonably sized CW-t test. For instance, with DGP 1 and a forecast horizon of 4, the NW version of the CW-t test compared against normal critical values has a rejection rate of 9.6% with $R = 40$, $\tilde{P} = 80$ and 9.3% with $R = 40$, $\tilde{P} = 120$.

For the same test, using the rectangular estimator of the standard deviation in the test statistic yields slightly better size performance. For example, in the DGP 1 experiment with a forecast horizon of 8 periods and $R = 120$, $\tilde{P} = 40$, the rejection rate of the CW-t test based on the rectangular estimator is 17.2%, while the rejection rate of the test based on the NW estimator is 18.3%. But it remains the case that the test can be significantly oversized, especially with small forecast samples, long horizons, and an alternative model with $k_w > 1$.

As with the NW estimator, using the West estimator of the standard deviation in the CW-t test often yields far too high a rejection rate, particularly with small $\tilde{P}$. Overall, the test based on the West estimator fares comparably — sometimes better, sometimes worse — to the test based on the NW estimator. For instance, with DGP 2 and a forecast horizon of 8 periods, using the West estimator yields a rejection rate of 32.7% with $R = 80$, $\tilde{P} = 20$ and 13.8% with $R = 80$, $\tilde{P} = 80$, compared to corresponding rejection rates of 27.7% and 15.5% based on the NW estimator.

Size performance is considerably better when the CW-t test is computed with the QS and HLN estimators (recall that the HLN estimator uses the rectangular variance estimate

and a finite-sample adjustment of the variance and test statistic). Once again, size tends to be an increasing function of the forecast horizon and alternative model size and a decreasing function of forecast sample size. For forecast samples of 40 or more observations, using the QS estimator often yields size below 10%. For example, with DGP 1, $R = 40$, $\tilde{P} = 120$, the rejection rate is 6.7% for the 4-step forecast horizon and 7.1% for the 8-step horizon. By reducing the forecast sample to $\tilde{P} = 80$ and moving to the larger alternative model of DGP 2, we raise the rejection rate to 10.0%. The QS-based test becomes over-sized — but to a much smaller degree than in the NW, rectangular, and West-based tests — in very small forecast samples ($\tilde{P} = 20$). For example, with DGP 2, a forecast horizon of 8 periods, and $R = 80$, $\tilde{P} = 20$, using the QS estimator with the CW-t test yields size of 15.9%. By comparison, the HLN-based test is less prone to being undersized, but a little more prone to being oversized in small samples (more so the longer the forecast horizon). For instance, with DGP 1, $R = 40$, $\tilde{P} = 120$, the HLN-based rejection rate is 7.8% for the 4-step forecast horizon and 9.1% for the 8-step horizon, compared to corresponding rates of 6.7% and 7.1% for the QS-based test. With DGP 2, $R = 120$, $\tilde{P} = 40$, the HLN-based rejection rate is 11.2% for the 4-step forecast horizon and 15.7% for the 8-step horizon, compared to corresponding rates of 9.2% and 12.2% for the QS-based test. Whether either the QS and HLN estimators can be viewed as best depends on one's concern with modest undersizing of QS versus modest oversizing of HLN.

Finally, for the MSE-t test compared to standard normal critical values, both the HLN and QS estimators yield the systematic undersizing that should be expected based on population-level asymptotics. Across all experiments in Tables 9 and 10, the size of the QS-based MSE-t test ranges from 0.3% to 9.8%, and the size of the HLN-based test ranges from 0.4% to 8.8%. The other HAC estimators — NW, rectangular, and West — can yield over-sized tests, if the forecast sample is small or the forecast horizon long. For example, in experiments with DGP 1, a forecast horizon of 8 periods, and $R = 80$, $\tilde{P} = 20$, the MSE-t tests based on the NW, rectangular, and West estimators have size of 18.9%, 14.7%, and 25.4%, respectively. With the same settings but for a forecast sample size of $\tilde{P} = 80$, the tests are undersized as expected, with corresponding rejection rates of 6.0%, 6.1%, and 6.8%.

7.1.5 Results: Null of equal accuracy in the finite sample

We turn now to tests under the null hypothesis of equal accuracy in the finite sample, for which Tables 11 and 12 report results. The results for the FRBS based on the NW estimator (of the V matrix that helps determine the bootstrap DGP) are consistent with those of Clark and McCracken (2011a). With small samples and multi-step forecasts, the MSE-F and MSE-t tests compared against FRBS critical values are slightly to modestly oversized. The size distortion tends to rise as the forecast sample shrinks, the forecast horizon increases, and the number of additional variables in the larger forecasting model (k_w) increases. For example, based on the NW HAC estimator, with $R = \tilde{P} = 80$, the MSE-F test has rejection rates of 12.6% with DGP 1 and $\tau = 4$, 14.3% with DGP 1 and $\tau = 8$, 16.0% with DGP 2 and $\tau = 4$, and 15.8% with DGP 2 and $\tau = 8$ (recall that $k_w = 1$ in DGP 1 and $k_w = 3$ in DGP 2). The size distortions tend to be a little smaller with the MSE-t test than MSE-F (however, as shown in Clark and McCracken (2011a), the MSE-t test also has lower power than the MSE-F test). In the same example, the MSE-t test has rejection rates of 11.0% with DGP 1 and $\tau = 4$, 12.5% with DGP 1 and $\tau = 8$, 14.2% with DGP 2 and $\tau = 4$, and 14.2% with DGP 2 and $\tau = 8$.

Using the rectangular estimator slightly reduces the size distortions of the MSE-F and MSE-t tests, with more noticeable improvements in DGP 2 (larger k_w) than DGP 1 (smaller k_w). For instance, with $R = \tilde{P} = 80$ and a forecast horizon of $\tau = 8$, the size of the MSE-F test in DGP 1 experiments edges down from 14.3% under the NW estimator to 14.1% under the rectangular estimator. In corresponding DGP 2 experiments, the rejection rate for MSE-F falls from 15.8% to 14.6%. Again, size distortions are slightly smaller for the MSE-t test than the MSE-F test. Reflecting these patterns, in empirical applications with properties similar to those of our experiments, for a forecast horizon of 4 periods or less and an unrestricted forecasting model that has only one variable more than the benchmark, the rectangular estimator may be seen as sufficient for obtaining reasonably accurate inference with the MSE-t test.

Using the QS estimator of the V matrix needed to set parameters of the FRBS yields somewhat larger gains in size performance. For instance, with $R = \tilde{P} = 80$ and a forecast horizon of $\tau = 8$, the size of the MSE-F test in DGP 1 experiments falls from 14.3% under the NW estimator to 11.1% under the QS estimator; the size of the MSE-t test declines from 12.5% (NW) to 10.4% (QS). In corresponding DGP 2 experiments, the rejection rate

for MSE-F falls from 15.8% (NW) to 12.9% (QS), and the rejection rate for MSE-t declines from 14.2% (NW) to 12.0% (QS). At the forecast horizon of four periods, in larger samples of forecasts in DGP 1, using the QS estimator with the FRBS can yield slightly undersized tests. For example, in the DGP 1 experiment with $R = 120$, $\tilde{P} = 40$, and $\tau = 4$, the MSE-F test has size of 8.8% when the QS estimator is used in the bootstrap. Overall, in empirical applications with properties similar to those of our experiments, the QS estimator seems to deliver reasonably good size properties with an unrestricted forecasting model that has only one variable more than the benchmark.

Across all experiment settings, using the West estimator of the HAC variance of the bootstrap yields the best size performance. For instance, with $R = \tilde{P} = 80$ and a forecast horizon of $\tau = 8$, the size of the MSE-F test in DGP 1 experiments falls from 14.3% under the NW estimator to 10.1% under the West estimator; the size of the MSE-t test declines from 12.5% (NW) to 9.7% (West). In corresponding DGP 2 experiments, the rejection rate for MSE-F falls from 15.8% (NW) to 9.6% (West), and the rejection rate for MSE-t declines from 14.2% (NW) to 10.1% (West). While the QS estimator often fares about as well as the West estimator when $\tilde{P}$ is smaller than R, at longer forecast horizons the West estimator fares much better than the QS estimator when $\tilde{P}$ is larger than R. Consider some of the experiments with DGP 2 and a forecast horizon of $\tau = 8$. With $R = 120$, $\tilde{P} = 40$, the MSE-F test has size of 11.2% under the FRBS based on the QS estimator and 9.0% under the bootstrap based on the West estimator. But with $R = 40$, $\tilde{P} = 120$, the MSE-F test has size of 13.2% under the FRBS based on the QS estimator and 9.1% under the bootstrap based on the West estimator.

Finally, we consider the MSE-t test compared to standard normal critical values. The t-tests based on the NW, rectangular, and West HAC estimators are prone to significant over-sizing if the forecast sample is small or the forecast horizon long. For example, in experiments with DGP 1, a forecast horizon of 8 periods, and $R = 80$, $\tilde{P} = 20$, the MSE-t tests based on the NW, rectangular, and West estimators have size of 27.0% 21.6%, and 32.2%, respectively. With the same settings but for a forecast sample size of $\tilde{P} = 80$, the tests are just modestly over-sized, with corresponding rejection rates of 14.3%, 14.2%, and 14.1%. The size of the test is much more accurate with the QS and HLN estimators of the standard deviation in the test statistic. For instance, in the DGP 1 experiment for the 8-step ahead horizon, with $R = 80$, $\tilde{P} = 80$, using the QS and HLN estimators yields

rejection rates of 9.6% and 12.2%, respectively, compared to rates of more than 14% for the NW, rectangular, and West estimators. Whether either the QS and HLN estimators can be viewed as best depends on one's concern with the tendency of QS to be undersized (more so than HLN) in some settings versus the tendency of HLN to be oversized (more so than QS) in other settings.

7.1.6 Results summary

Based on these results, we can offer some recommendations for obtaining accurate inference in tests applied to multi-step forecasts from nested models, taking as given a desire to keep variance computations as simple as possible. While other estimators can work in more limited conditions (e.g., forecast horizons that aren't too long and forecast samples that are fairly large), the following seem to work well in general conditions.

- Tests of equal accuracy in population compared against critical values obtained with the no-predictability fixed regressor bootstrap of Clark and McCracken (2011b): simply use the Newey and West (1987) estimator in computing test statistics.

- Tests of equal accuracy in population compared against standard normal critical values: use either the pre-whitened quadratic spectral estimator of Andrews and Monahan (1992) or the adjusted variance developed in Harvey, Leybourne, and Newbold (1997) in computing the MSE-t and CW-t tests (this will yield a CW-t test with empirical size about equal to nominal and a MSE-t test that doesn't yield spurious rejections with small samples and long horizons).

- Tests of equal accuracy in the finite sample compared against critical values obtained with the fixed regressor bootstrap of Clark and McCracken (2011a): use the HAC estimator of West (1997) to compute the V matrix that helps determine the bootstrap parameterization, and use the Newey and West (1987) estimator in computing the denominators of t-tests.

- Tests of equal accuracy in the finite sample compared against standard normal critical values: use either the pre-whitened quadratic spectral estimator of Andrews and Monahan (1992) or the adjusted variance developed in Harvey, Leybourne, and Newbold (1997) in computing the MSE-t test.

7.2 Size corrections

As with any testing that is based upon asymptotic approximations, there is always the concern that the asymptotic distribution does not match well with the finite sample distribution of the test statistic. That is, while it may be the case that a t-type test of zero mean prediction error of the form

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(\hat{u}_{t+\tau}-0)/\hat{\Omega}^{1/2} \tag{35}$$

is asymptotically standard normal, it may not be the case that the standard normal approximation works well in a sample of size (say) $T=100$ with $P=R=50$.

In this section we highlight a particular type of size-correction mechanism suggested in Giacomini and Rossi (2009) that is based on an extension of the theory in West (1996). To understand the source of their proposed size correction, note that the theory developed in West (1996) is based upon a particular decomposition of the moment condition $P^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma)$:

$$\begin{aligned}(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma) &= (P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\beta^*)-\gamma) \\ &+ FB(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}H(t)+o_p(1).\end{aligned} \tag{36}$$

The first right-hand side component captures the part of the test statistic that would exist if the parameters were known and did not need to be estimated. The second component captures the effect of parameter estimation error on the test statistic. Each of these two components can be asymptotically normal and hence when added together, the term $(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma)$ is asymptotically normal with an asymptotic variance that is, in general, affected by each of the two subcomponents.

The size correction proposed by Giacomini and Rossi (2009) arises not from either of these two terms but rather from a judicious decomposition of the residual term $o_p(1)$ in equation (36). They note that while it is certainly true that this residual component is asymptotically irrelevant, it might be the case that at least part of it is important in finite samples. Their proposed size correction is based on a modest extension of equation (36)

that is based on the second order term in a Taylor expansion:

$$\begin{aligned}(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma) &= (P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\beta^*)-\gamma) \\ &+ FB(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}H(t) \\ &+ 0.5((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}H'(t)B'(E\frac{\partial^2 f_{t+\tau}(\beta^*)}{\partial\beta\partial\beta'})BH(t))+o_p(1).\end{aligned} \tag{37}$$

As shown in West (1996), $((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}H'(t)B'(E\frac{\partial^2 f_{t+\tau}(\beta^*)}{\partial\beta\partial\beta'})BH(t))$ is $o_p(1)$. That said, in finite samples this term might be sufficiently large to prevent the test statistic from being well approximated by a standard normal distribution. Giacomini and Rossi (2009) therefore suggest a size-corrected form of the test statistic that subtracts an estimate of the mean of the second order term and then bases inference on the standard normal distribution. Specifically they recommend using a size-corrected version of the test statistic that takes the form

$$((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma)-SC_T)/\hat{\Omega}^{1/2}, \tag{38}$$

where the size-correcting term SC_T takes a form that depends upon the sampling scheme being used:

$$\text{fixed and rolling}: T^{-1/2}0.5(\hat{\pi})^{1/2}(1+\hat{\pi})^{1/2}tr\left(\hat{B}'\left((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\frac{\partial^2 f_{t+\tau}(\hat{\beta}_t)}{\partial\beta\partial\beta'}\right)\hat{B}\hat{S}_{hh}\right) \tag{39}$$

$$\text{recursive}: -T^{-1}0.5\left(1+\hat{\pi}^{-1}\right)^{1/2}\ln\left(1+\hat{\pi}\right)tr\left(\hat{B}'\left((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\frac{\partial^2 f_{t+\tau}(\hat{\beta}_t)}{\partial\beta\partial\beta'}\right)\hat{B}\hat{S}_{hh}\right). \tag{40}$$

This derivation yields two broad conclusions.

1. Holding $tr\left(\hat{B}'\left((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\frac{\partial^2 f_{t+\tau}(\hat{\beta}_t)}{\partial\beta\partial\beta'}\right)\hat{B}\hat{S}_{hh}\right)$ constant, the size-correcting term is larger for the fixed and rolling schemes than the recursive. This occurs since for all $\hat{\pi}$, $\hat{\pi}^{1/2}(1+\hat{\pi})^{1/2}$ is larger than $(1+\hat{\pi}^{-1})^{1/2}\ln(1+\hat{\pi})$.

2. Holding $tr\left(\hat{B}'\left((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\frac{\partial^2 f_{t+\tau}(\hat{\beta}_t)}{\partial\beta\partial\beta'}\right)\hat{B}\hat{S}_{hh}\right)$ constant, the size-correcting term is increasing in $\hat{\pi}$ for all sampling schemes. Hence one expects that the size-correction will be most useful when the initial estimation sample size R is small relative to the total

sample size T. Accordingly, size correction may become important if P/R is set high to achieve high power (in light of the evidence above that, in many settings, power is maximized by making P/R large).

8 On the Choice of Sample Split

In any out-of-sample testing environment one has to decide how to split the sample into in-sample and out-of-sample portions. That is, if one has access to observables from $t = 1, \ldots, T$, in order to conduct a pseudo-out-of-sample forecasting exercise one has to determine how much data to withhold for the initial estimation sample (R) and how much to use for forecast evaluation (P). In this section we provide some tentative guidance towards making that decision when the goal is to maximize power.

We separate our analysis into three distinct parts. First we provide some tentative guidance when asymptotic inference follows from the results in West (1996) — and hence notably is valid for comparisons of non-nested models. We then provide some discussion for nested model comparisons based on recent work by Hansen and Timmermann (2011). Finally, we discuss recent work by both Hansen and Timmermann (2011) and Rossi and Inoue (2011) on methods for conducting inference that avoids the sample-split issue all together. Throughout we focus exclusively on tests of population-level predictive ability.

8.1 Optimality in the West (1996) framework

Recall from section 3.1.1 that West (1996) shows that under the null hypothesis $H_0 : Ef_{t+\tau} = \gamma$, a test statistic of the form

$$(P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} (f_{t+\tau}(\hat{\beta}_t) - \gamma)/\hat{\Omega}^{1/2} \tag{41}$$

can be asymptotically standard normal if estimation error is appropriately accounted for when constructing $\hat{\Omega}$. Suppose that instead of the null hypothesis holding, there exists a sequence of local alternatives satisfying $Ef_{t+\tau} = \gamma + T^{-1/2}\delta$. In this environment it is straightforward to show that

$$(P - \tau + 1)^{-1/2} \sum_{t=R}^{T-\tau} (f_{t+\tau}(\hat{\beta}_t) - \gamma)/\hat{\Omega}^{1/2} \rightarrow^d N(\left(\frac{\pi}{1+\pi}\right)^{1/2} \left(\frac{\delta}{\Omega^{1/2}}\right), 1), \tag{42}$$

which is asymptotically normal with unit variance but has a non-zero asymptotic mean that depends explicitly on the sample-split parameter π through both $\frac{\pi}{1+\pi}$ and Ω. In practice

this type of test is typically two-sided and hence rather than work with the statistic in (36) we look at its square. Under the sequence of local alternatives we immediately have that

$$\left((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-\gamma)/\hat{\Omega}^{1/2}\right)^2 \rightarrow^d \chi^2(1;\Lambda), \tag{43}$$

a non-central χ^2 variate with a non-centrality parameter Λ that varies with the estimation scheme because Ω varies with the estimation scheme:

$$\text{Fixed}, 0 \leq \pi < \infty : \Lambda = (\frac{\pi}{1+\pi})(\frac{\delta^2}{S_{ff}+\pi FBS_{hh}B'F'}) \tag{44}$$

$$\text{Rolling}, 0 \leq \pi \leq 1 : \Lambda = (\frac{\pi}{1+\pi})(\frac{\delta^2}{S_{ff}+(\pi)FBS'_{fh}+(\pi-\frac{\pi^2}{3})FBS_{hh}B'F'}) \tag{45}$$

$$\text{Rolling}, 1 \leq \pi < \infty : \Lambda = (\frac{\pi}{1+\pi})(\frac{\delta^2}{S_{ff}+(2-\frac{1}{\pi})FBS'_{fh}+(1-\frac{1}{3\pi})FBS_{hh}B'F'}) \tag{46}$$

$$\text{Recursive}, 0 \leq \pi \leq \infty : \Lambda = (\frac{\pi}{1+\pi})(\frac{\delta^2}{S_{ff}+2(1-\frac{1}{\pi}\ln(1+\pi))(FBS'_{fh}+FBS_{hh}B'F')}). \tag{47}$$

Maximizing power is then equivalent to choosing the value of π that maximizes the non-centrality parameter associated with the sampling scheme being used. Doing so we obtain

$$\text{Fixed } \pi^* = \left\{ \begin{array}{ll} \infty & \text{if } F=0 \\ \left(\frac{S_{ff}}{FBS_{hh}B'F'}\right)^{1/2} & \text{else} \end{array} \right\} \tag{48}$$

$$\text{Rolling } \pi^* = \infty \tag{49}$$

$$\text{Recursive } \pi^* = \left\{ \begin{array}{ll} \infty & \text{if } F=0 \text{ or } -FBS'_{fh} = FBS_{hh}B'F' \\ \left(\frac{2+\pi^*}{\pi^*}\right)\ln(1+\pi^*) = 2+\frac{1}{2}\left(\frac{S_{ff}}{FBS'_{fh}+FBS_{hh}B'F'}\right) & \text{else} \end{array} \right\} \tag{50}$$

From these derivations, we can draw the following conclusions.

1. In each case, when $F=0$ we find that the optimal sample split is one that chooses the ratio P/R to be large. Perhaps the most important application of this optimality result

is in cases in which two OLS-estimated non-nested models are being compared based on their mean square errors. However, one should note that strictly speaking the $\pi^* = \infty$ case cannot literally be taken to be true for the fixed and rolling schemes since the results in West (1996) only apply when $0 \leq \pi < \infty$. Even so, when $F = 0$ it is clearly the case that the non-centrality parameter is monotone increasing in π and hence the optimal value of π is arbitrarily large.

2. For both the rolling and recursive schemes, in those cases for which $-FBS'_{fh} = FBS_{hh}B'F'$, we find that the optimal sample split is one that chooses the ratio P/R to be large. While this case may seem an unlikely coincidence, West (1996) and West and McCracken (1998) show that this happens fairly easily when evaluating OLS-estimated linear models using tests of zero-mean prediction error or efficiency when the model errors are conditionally homoskedastic and serially uncorrelated.

3. When estimation error is not asymptotically irrelevant and hence $\Omega \neq S_{ff}$, the optimal sample split can take values that are not arbitrarily large and in fact can be quite small depending on the covariance structure of the observables. One simple example occurs in the cases described in point 2 but when the fixed scheme is used: when evaluating OLS-estimated linear models using tests of zero-mean prediction error or efficiency it can be the case that $-FBS'_{fh} = FBS_{hh}B'F' = S_{ff}$ and hence we find that the optimal sample split uses half of the observables to estimate the model parameters and the other half to evaluate the forecasts.

When $\Omega \neq S_{ff}$, the optimal sample split is more difficult to interpret for the recursive scheme, for which there does not seem to be a closed form solution. Rather, the optimal sample split must be inferred numerically given values of S_{ff}, FBS'_{fh}, and $FBS_{hh}B'F'$.

4. In general, when estimation error is asymptotically relevant the optimal sample split is finite but depends on unknown nuisance parameters. Using the methods described in section 3.1.1, these parameters can be estimated using the observables and hence one can imagine constructing a feasible variant of the optimal sample split parameter π^*. Of course taking such an approach precludes the optimal sample split since it is very unlikely that in any finite sample the estimate will match the optimal value. Even worse, estimating the optimal sample split parameter requires conducting a preliminary pseudo out-of-sample exercise which by its vary nature constitutes pre-testing. Thus any out-of-sample inference based on an estimated optimal sample split is unlikely to match the theory for which it

was designed. Put more bluntly, if we let $\hat{\pi}^*$ denote the estimated optimal sample split parameter, $\hat{R}^* = [T\frac{1}{1+\hat{\pi}^*}]$, and $\hat{P}^* = T - \hat{R}^* + \tau$, it is not obvious that the statistic

$$(\hat{P}^* - \tau + 1)^{-1/2} \sum_{t=\hat{R}^*}^{T-\tau} (f_{t+\tau}(\hat{\beta}_t) - \gamma)/\hat{\Omega}^{*1/2} \tag{51}$$

is asymptotically standard normal.

5. While not a proof, based upon the analytics above it seems reasonable to suggest a simple rule of thumb: when choosing a sample split one should choose a value of P/R that is at least 1 and perhaps much higher. To be clear, this argument is based solely on a desire to maximize power and not to reduce any potential size distortions. For example, as we saw in section 7.1.4, we are more likely to observe finite sample size distortions when P/R is large, especially when the fixed or rolling schemes are being used. Fortunately, as shown in section 7.2, for non-nested models a simple size correction mechanism is easily introduced to the test statistic that helps ameliorate the issue.

8.2 Optimality for nested model comparisons

As noted in Clark and McCracken (2001), among others, the analytics in West (1996) do not apply when constructing either tests of equal MSE or tests of encompassing for two models that are nested under the null. As such, the analytics related to the optimal choice of sample split cannot be inferred from the results described in the previous section. Regardless, Hansen and Timmermann (2011) present results that are quite similar in the sense that the optimal sample split is one that chooses the ratio P/R to be large.

Consider the case discussed in section 3.1.2 where two nested OLS-estimated linear models are being compared, such that model 2 nests model 1 and hence $\beta_2^* = (\beta_1^{*\prime}, \beta_w^{*\prime})' = (\beta_1^{*\prime}, 0)'$ under the null. But as we did for the results above, suppose that, instead of the null hypothesis holding, there exists a sequence of local alternatives satisfying $\beta_{2,T} = (\beta_1', T^{-1/2}\beta_w')'$. In section 3.2.2 we showed that under the recursive scheme we obtain[36]

$$\text{MSE-}F \rightarrow^d \{2\Gamma_1 - \Gamma_2\} + 2\{\Gamma_4\} + \{\Gamma_5\}. \tag{52}$$

Inoue and Kilian (2004) obtained a similar result, in a slightly less general model setup, in a comparison of the power of in-sample and out-of-sample tests of population-level predictive ability.

[36] Results for the rolling and fixed schemes are similar.

In equation (52) we see that the sequence of local alternatives only affects the asymptotic distribution through Γ_4 and Γ_5. Moreover, it is fairly intuitive to interpret $\Gamma_5 = (1 - \lambda)\beta_w' F_2^{-1}\beta_w/\sigma^2$ as the non-centrality parameter of the asymptotic distribution in the same way as we did above for Λ in the West-based analytics. If we treat this term as the objective function and maximize it with respect to π we quickly find that the optimal value of the sample split is one that chooses the ratio P/R to be large. The analytical argument presented here reinforces the simulation-based evidence provided in Clark and McCracken (2001, 2005a) and McCracken (2007). A more formal discussion of the optimal sample split is given in Hansen and Timmermann (2011).

8.3 Sample-split robust methods

Motivated at least in part by the potential for sensitivity of forecast evaluation results to sample choice, Hansen and Timmermann (2011) and Rossi and Inoue (2012) develop methods for testing the null of equal predictive ability across different sample splits. In the former, the null is equal predictive ability at the population level; the latter considers equal predictive ability at the population level and in the finite sample. One concern is with the effects of data mining: in practice, one might search across sample splits (or be influenced by results in other studies) for a test result that appears significant, without taking the search into account in gauging significance. The other concern is with power: as noted above, some sample splits might yield greater power than others. In light of these concerns, tests that explicitly consider a range of samples might have advantages.

In these studies, it is assumed that, in a given data set, forecasts are evaluated over a range of sample splits. More specifically, continuing to let R denote the last observation used in estimation for forming the first forecast, forecast tests may be formed using R settings of between R_l and R_u. Under this multiple-sample approach, one might consider the maximum of the sequence of test statistics computed for a range of samples. For example, with the MSE-F test, the robust test would take the form

$$\max_{R=R_l,\ldots,R_u} \text{MSE-}F(R) = \max_{R=R_l,\ldots,R_u} \left(\sum_{t=R}^{T-\tau} \hat{d}_{t+\tau}\right) / \hat{\sigma}_2^2(R),$$

where $\hat{\sigma}_2^2(R)$ denotes the MSE of model 2 for the sample split at observation R.

Focusing on nested models, Hansen and Timmermann (2011) use the asymptotic framework of Clark and McCracken (2001, 2005a) and McCracken (2007) to develop the asymptotic distribution of the maximum of the MSE-F test. As detailed below, Monte Carlo

simulations confirm that searching across samples without taking the search into account yields spurious findings of predictive ability. For 1-step ahead forecasts (with conditional homoskedasticity), Hansen and Timmermann consider a local alternative (drawing on their results that simplify the asymptotic distribution of McCracken (2007)) to assess power, which indicates that power rises as the forecast sample grows — a finding consistent with our analysis in the preceding section. Out of concern that the marginal distribution of the test statistic computed for each sample split varies with the sample split, Hansen and Timmermann propose using nominal p-values for each individual sample split instead of test statistics for each split. More specifically, they propose comparing the minimum p-value with critical values obtained by Monte Carlo simulations of an asymptotic distribution (given in the paper) that is a functional of Brownian motion.

Rossi and Inoue (2011) develop results for both non-nested and nested models. With non-nested models, Rossi and Inoue use high-level assumptions that rest on the asymptotic framework of West (1996). They consider two test statistics, one that averages a normalized loss differential across different sample splits and the other that is the maximum of the normalized loss differential across sample splits, where the sample is split at each possible observation between R_l and R_u:

$$\begin{aligned} \text{sup test} &= \sup_{R=R_l,\ldots,R_u} \frac{1}{\hat{\sigma}_R} T^{-1/2} \bar{d}(R) \\ \text{average test} &= \frac{1}{R_u - R_l + 1} \sum_{R=R_l}^{R_u} |\frac{1}{\hat{\sigma}_R} T^{-1/2} \bar{d}(R)|, \end{aligned}$$

where $\bar{d}(R)$ denotes the average loss differential for the forecast sample that begins with observation $R + \tau - 1$ and $\hat{\sigma}_R^2$ denotes a consistent estimate of the long-run variance of the loss differential for the same sample. The null hypothesis is that, in population, the average loss differential is 0 for all sample splits (all R considered).

Under West-type conditions that imply the partial sum of the loss differential obeys a functional central limit theorem, Rossi and Inoue show that the null asymptotic distributions of the test statistics are functions of (univariate) standard Brownian motion. The distributions depend on the sample fractions R_l/T and R_u/T but no other parameters. Rossi and Inoue provide a table of asymptotic critical values obtained by Monte Carlo simulation.

For nested models, Rossi and Inoue (2011) provide results for the F-type test of forecast encompassing developed in Clark and McCracken (2001), denoted ENC-F above. In this

case, Rossi and Inoue rely on the asymptotics of Clark and McCracken (2001) and show that, for 1-step ahead, conditionally homoskedastic forecast errors, the asymptotic distribution for the average and maximum of the statistic across sample splits is also a function of standard Brownian motion, with dependence on the range of sample splits and the number of additional parameters in the larger model. Again, Rossi and Inoue use Monte Carlo simulations of the asymptotic distribution to obtain critical values, provided in tables in the paper.

For the case of estimation samples that can be viewed as small relative to the forecasting sample, Rossi and Inoue (2011) develop multiple-sample tests based on the Clark and West (2006, 2007) t-test for equality of adjusted MSEs. They propose two tests — one a maximum and the other an average — robust to multiple samples. For example, the maximum version takes the form

$$\text{sup test} = \sup_{R=R_l,\ldots,R_u} \frac{1}{\hat{\sigma}_R} T^{-1/2} \overline{cw}(R), \tag{53}$$

where $\overline{cw}(R)$ denotes the average Clark-West loss differential for the forecast sample that begins with observation $R + \tau - 1$ and $\hat{\sigma}_R$ denotes a consistent estimate of the long-run variance of the loss differential for the same sample. The null hypothesis is that, in population, the average Clark-West loss differential is 0 for all sample splits (all R considered). In this case, too, the null asymptotic distributions of the test statistics are functions of (univariate) standard Brownian motion, with critical values available from tables provided by the authors.

Finally, Rossi and Inoue (2011) also develop multiple sample-robust versions of a range of regression-based tests of predictive ability, including tests for bias, efficiency, the Chong and Hendry (1986) form of encompassing, and serial correlation. Under the assumption that the partial sum of a loss function obeys a functional central limit theorem, Rossi and Inoue show that the maximum and average of Wald tests formed for a range of sample splits have limiting distributions that are functions of Brownian motion, depending on only the sample fractions R_l/T and R_u/T . These results will apply under the conditions described in West (1996) and West and McCracken (1998) that are necessary to obtain standard distributions for tests applied to a single forecast sample; in many cases, the relevant variance matrix will need to be computed to account for the effects of parameter estimation error.

Monte Carlo evidence in Hansen and Timmermann (2011) and Rossi and Inoue (2011) shows that searching across sample splits without accounting for it in inference can yield

material size distortions. However, in both studies, the presumed searches are extensive, across many different (continuous) sample splits. In practice, researchers probably engage in more limited searches, checking just a few (discrete) sample splits. The impacts of more limited searches are more modest. At any rate, Monte Carlo experiments in Rossi and Inoue (2011) also indicate that their proposed tests have reasonable size properties. As to power in the finite sample, Rossi and Inoue (2011) present Monte Carlo evidence that using their tests can offer important gains in power over the approach of conducting a test for a single split. However, it seems that most of the power gains come with instabilities in the data generating process and forecasting models. For example, if the predictive content of one variable for another fell 3/4 of the way through the data sample, searching for predictive content across a wide range of samples increases the chances of detecting predictive content relative to the chance of finding the content with a test based on one short forecast sample based on, say, just the last 1/4 of the sample.

9 Why Do Out-of-Sample Forecast Evaluation?

As indicated in the Introduction, forecast evaluation has long been an important tool for evaluating models. While modern usage seems to have picked up since the pioneering work of Fair and Shiller (1989, 1990) and Meese and Rogoff (1983, 1988), West (2006) observes that Wilson (1934) represents an early example of a long tradition of using predictive ability to assess models.

This common reliance on forecast evaluation likely reflects several considerations. First, many individuals and institutions (such as central banks) have need of out-of-sample forecasts. In these cases, forecast evaluation is intended to be a useful tool for assessing past performance and gauging the potential for future effectiveness — for example, identifying the model that has been best in the past for the purpose of using it to forecast going forward, in the hope of forecasting as accurately as possible in the future. Second, for some practitioners and researchers, forecast evaluation is viewed as useful for guarding against structural instabilities and model overfitting. By now, based on evidence in studies such as Stock and Watson (1996, 2003), many empirical relationships are thought to be unstable over time. In light of the common finding that that in-sample predictive ability fails to translate into out-of-sample predictive ability (e.g., Stock and Watson 2003, Goyal and Welch 2008), out-of-sample forecast comparisons may be useful for avoiding models that

are unstable.

As to overfitting, it is widely believed that empirical modeling is prone to overfitting (see, for example, Ashley, Granger, and Schmalensee (1980), Chatfield (1995), Leamer (1978), Lo and MacKinlay (1990), and Lovell (1983)). In particular, various forms of data mining may lead a researcher to falsely conclude that some variable x has explanatory power for another variable y. As discussed by Hoover and Perez (1999) and Lovell (1983), the data mining may take the form of a search across candidate models for y. For example, a researcher might search across 10 different x variables to find the one that has the most explanatory power for y. The data mining may also more generally reflect the results of a profession–wide search that has affected the set of candidate variables, a possibility noted by West (1996) and considered in some detail by Denton (1985) and Lo and MacKinlay (1990).

The hope of reducing the probability of overfitting appears to lead some researchers to examine out–of–sample forecasts for evidence of predictive power. In the simplest case, if in–sample evidence suggests some x has explanatory power for y, a researcher may construct competing forecasts of y, using one model of y that includes x and another that does not. If x truly has explanatory power for y, forecasts from the model including x should be superior. Accordingly, Ashley, Granger, and Schmalensee (1980) advocate using out–of–sample forecast comparisons to test Granger causality.

Notwithstanding these rationales and the large literature on forecast evaluation, the question of why one should conduct out-of-sample analysis has remained a source of some controversy. Some studies explicitly steer away from the question by simply taking the interest in forecasts as given: for example, Hubrich and West (2010) report "...we do not attempt to explain or defend the use of out-of-sample analysis. As is usual in out-of-sample analysis, our null is one that could be tested by in-sample tools.... Our aim is not to argue for out-of-sample analysis, but to supply tools to researchers who have concluded that out-of-sample analysis is informative for the application at hand."

Of the various rationales for forecast evaluation, the intention of evaluating the forecasts to assess the models for their actual value in forecasting should be the least controversial. If one's goal is to use a model for out-of-sample forecasting, it seems reasonable to use historical forecast performance to judge the model. Logically, for this line of reasoning, the challenge is that, with nested forecasting models, many of the existing testing methods — for population-level predictive ability — are equivalent to testing exclusion restrictions on

the larger forecasting model. Of course, as emphasized in Inoue and Kilian (2004), these same restrictions could be tested with conventional in-sample methods (e.g., conventional Wald tests), which will often have better power than the available forecast-based tests. The development of methods for testing equal accuracy in the finite sample (by Giacomini and White (2006), Clark and McCracken (2011a), and Calhoun (2011)) can help to ameliorate this concern. As described in section 3.2, these tests address predictive ability in a finite sample, which seems closer to the question of focus for those interested in actual value in forecast models. In this case, at a minimum, tests for predictive ability in population can have value as first-pass screens. With a test for finite-sample predictive ability representing a higher bar than a test for population-level predictive ability, if a population-level comparison doesn't indicate a larger model is better than a smaller model, neither will a finite-sample comparison.

The value of forecast-based tests for avoiding instabilities and overfitting remains somewhat more controversial, although we would argue there can indeed be important value. For picking up instabilities, Clark and McCracken (2005b) show (with asymptotic theory and Monte Carlo evidence) that in-sample explanatory power is readily found because the usual F-test indicates Granger causality or predictive ability if it existed at any point in the sample. Out-of-sample predictive power can be harder to find because the results of out-of-sample tests are highly dependent on the timing of the predictive ability — whether the predictive ability existed at the beginning or end of the sample, and where a break occurred relative to the start of the forecast sample. Overall, out-of-sample tests are effective at revealing whether one variable has predictive power for another at the end of the sample. More recently, Inoue and Rossi (2005) and Giacomini and Rossi (2009) have developed a variety of tools for detecting breakdowns in predictive content.

As to overfitting, Monte Carlo evidence in Clark (2004) confirms what may be interpreted as the original logic of Ashley, Granger, and Schmalensee (1980). If a researcher uses a given data sample to search across model specifications, the resulting model is likely to be overfit. However, evaluating forecasts in a subsequent sample that was not part of the sample used to determine the model specification is not subject to distortions, in the sense that the forecast-based tests are correctly sized. Be that as it may, Kilian and Inoue (2004) emphasize that the out-of-sample analysis can also be subject to data mining. If a researcher also searches across forecast model performance, both out-of-sample and in-sample

inference are subject to overfitting (size distortions). In this case, out-of-sample tests have no advantage over in-sample tests, and can be at a power disadvantage. That said, the recently developed methods for evaluating multiple forecasting models (reviewed in section 5) and evaluating forecasts across multiple sample splits (reviewed in section 8) provide additional tools for ensuring that forecast-based inferences avoid contamination from data mining.

10 Conclusion

Taking West's (2006) survey as a starting point, this paper reviews recent developments in the evaluation of point forecasts. To put recent work in a broader context, we begin by briefly covering the state of the literature as of the time of West's writing. Our chapter extends West's overview for practitioners by including a brief exposition of the derivations of some of the key results in the literature. The bulk of the chapter focuses on recent developments, including advancements in the evaluation of forecasts at the population level (based on true, unknown model coefficients), the evaluation of forecasts in the finite sample (based on estimated model coefficients), and the evaluation of conditional versus unconditional forecasts.

In this chapter, we also hone in on two outstanding issues in the literature, and present some original results on these issues. The first is obtaining accurate inference in evaluation of finite samples of multi-step forecasts. The second issue is the optimization of power in determining the split of a sample into in-sample and out-of-sample portions. We provide a Monte Carlo assessment of options — alternative estimators of heteroskedasticity-and-autocorrelation (HAC) consistent variances — for obtaining finite sample inferences more reliable than those evident from some prior Monte Carlo work. We also present some original analysis extending West's (1996) results to include conditional forecasts.

11 Appendix: Asymptotic Derivations for Out-of-Sample Inference: Examples

In this chapter we have provided an overview of recent developments in forecast evaluation with an emphasis on how to conduct inference in a variety of applications. One thing we have purposefully avoided is the detailed mathematics behind most of the results. In this section we take a middle ground and provide some simple examples of how the asymptotic theory is derived.

In the first two subsections we provide step-by-step guides as to how the analytics work when we follow the style of proof used in West (1996) and Clark and McCracken (2001), where both P and R are allowed to diverge with the total sample size T. In the final subsection we follow the style of proof used in Giacomini and White (2006), where P is allowed to diverge with the total sample size T but R is a finite constant. To make the presentation as clear as possible, in the first two sections we focus exclusively on the fixed scheme and hence $\hat{\beta}_t = \hat{\beta}_R$, while in the final section we use the rolling scheme.

11.1 Test of zero mean prediction error: West (1996)

Suppose we are forecasting with a linear OLS-estimated regression model of the form $y_{t+1} = x_t'\beta^* + u_{t+1}$, where the vector of predictors contains an intercept and hence the first element of x_t is 1. Using this model, a sequence of 1-step ahead forecast errors $\hat{u}_{t+1} = y_{t+1} - x_t'\hat{\beta}_R$ are constructed. Based on these forecast errors we wish to test the null hypothesis H_0 : $E(u_{t+1}) = 0$ for all t. To do so we follow the analytics of West (1996) and base our statistic on the scaled out-of-sample average of the forecast errors $P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1}$. To derive the asymptotic distribution of this scaled average note that

$$\begin{aligned} P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1} &= P^{-1/2}\sum_{t=R}^{T-1}(y_{t+1} - x_t'\hat{\beta}_R) = P^{-1/2}\sum_{t=R}^{T-1}(y_{t+1} - x_t'\beta^*) - P^{-1/2}\sum_{t=R}^{T-1}x_t'(\hat{\beta}_R - \beta^*) \\ &= P^{-1/2}\sum_{t=R}^{T-1}u_{t+1} + (\frac{P}{R})^{1/2}(-P^{-1}\sum_{t=R}^{T-1}x_t')(R^{1/2}(\hat{\beta}_R - \beta^*)) \qquad (54) \\ &= P^{-1/2}\sum_{t=R}^{T-1}u_{t+1} + (\frac{P}{R})^{1/2}(-P^{-1}\sum_{t=R}^{T-1}x_t')(R^{-1}\sum_{s=1}^{R-1}x_s x_s')^{-1}(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_s). \end{aligned}$$

So far we have used only algebra. In order to derive the asymptotic distribution of $P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1}$ we need to fall back on the assumptions in West (1996) loosely presented in section 3.1.1. Specifically we need to assume that the sequence $(u_{s+1}, x_s)'$ is covariance

stationary, mixing, and has bounded fourth moments. With these assumptions in hand it is clear that if we let both P and R diverge such that $\lim_{P,R\to\infty} P/R = \pi$, we obtain

$$P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1} = P^{-1/2}\sum_{t=R}^{T-1}u_{t+1} + \pi^{1/2}(-Ex_t')(Ex_sx_s')^{-1}(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_s) + o_p(1).$$

If we let both P and R tend to infinity, both $P^{-1/2}\sum_{t=R}^{T-1}u_{t+1}$ and $R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_s$ are asymptotically normal with zero mean and asymptotic variances S_{ff} and S_{hh}, respectively. Since a linear combination of normal random variates is normally distributed we immediately find that

$$P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1} \to^d N(0,\Omega), \tag{55}$$

with

$$\begin{aligned}\Omega &= S_{ff} + \pi(-Ex_t')(Ex_sx_s')^{-1}S_{hh}(Ex_sx_s')^{-1}(-Ex_t) \\ &= S_{ff} + \pi(Ex_t')(Ex_sx_s')^{-1}S_{hh}(Ex_sx_s')^{-1}(Ex_t),\end{aligned} \tag{56}$$

which matches exactly with the formula for Ω under the fixed scheme in equation (2) of section 3.1.1.

The formula for Ω simplifies even further if we are willing to assume that the errors u_{t+1} are serially uncorrelated and conditionally homoskedastic. If this is the case we know that $S_{ff} = \sigma^2$ and $S_{hh} = \sigma^2 Ex_sx_s'$. Moreover, if we note that since the first element of x_t is 1, we have $(Ex_t')(Ex_sx_s')^{-1} = (1, 0')$, and hence

$$\begin{aligned}\Omega &= \sigma^2 + \pi\sigma^2(Ex_t')(Ex_sx_s')^{-1}(Ex_t) \\ &= \sigma^2(1+\pi).\end{aligned} \tag{57}$$

In this special case an asymptotically valid test of zero mean prediction error is constructed as

$$\frac{P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1}}{\sqrt{(1+(\frac{P}{R}))(P^{-1}\sum_{t=R}^{T-1}\hat{u}_{t+1}^2)}} \tag{58}$$

and inference can be conducted using standard normal critical values.

This last statistic also provides a simple foil for giving intuition on how the sample split-robust asymptotics in Rossi and Inoue (2011) work when implemented using the fixed scheme. For example, suppose we construct this statistic for each $R_j = R_l, ..., R_u$ satisfying $R_j + P_j = T$. Their proposed statistic takes the form

$$\sup_{R=R_l,\ldots,R_u}\frac{P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1}}{\sqrt{(1+(\frac{P}{R}))(P^{-1}\sum_{t=R}^{T-1}\hat{u}_{t+1}^2)}}. \tag{59}$$

The statistic is not asymptotically normal but is instead the supremum of a Gaussian process for which critical values can be simulated. Interestingly, this specific statistic is very closely related to one designed by Wright (1997) in the context of tests for structural change.

11.2 Test of equal predictive ability for nested models: Clark and McCracken (2001)

Suppose we are forecasting with two linear OLS-estimated regression models of the form $y_{t+1} = x'_{i,t}\beta_i^* + u_{i,t+1}$, where the vector of predictors $x_{2,t}$ contains the predictors in model 1 as well as an additional set of predictors $x_{w,t}$ and hence $x_{2,t} = (x'_{1,t}, x'_{w,t})'$. Using this model a sequence of 1-step ahead forecast errors $\hat{u}_{i,t+1} = y_{t+1} - x'_{i,t}\hat{\beta}_{i,R}$ are constructed. Again, to simplify exposition, we assume a fixed estimation scheme. Based on these forecast errors we wish to test the null hypothesis $H_0 : E(u_{1,t+1}^2 - u_{2,t+1}^2) = 0$ for all t. To do so we follow the analytics of Clark and McCracken (2001) and base our statistic on the scaled out-of-sample average of the loss differential $\sum_{t=R}^{T-1}\left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)$. To derive the asymptotic distribution of this scaled average note that

$$\begin{aligned}\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) &= \sum_{t=R}^{T-1}((y_{t+1} - x'_{1,t}\hat{\beta}_{1,R})^2 - (y_{t+1} - x'_{2,t}\hat{\beta}_{2,R})^2) \\ &= \sum_{t=R}^{T-1}((y_{t+1} - x'_{1,t}\beta_1^*) - x'_{1,t}(\hat{\beta}_{1,R} - \beta_1^*))^2 - ((y_{t+1} - x'_{2,t}\beta_2^*) - x'_{2,t}(\hat{\beta}_{2,R} - \beta_2^*))^2).\end{aligned}$$

This simplifies significantly since, under the null, $x'_{1,t}\beta_1^* = x'_{2,t}\beta_2^*$. If we square the terms

inside the summation and define $J = (I, 0)'$ and $u_{t+1} = u_{1,t+1} = y_{t+1} - x'_{1,t}\beta_1^*$ we obtain

$$\begin{aligned}
&\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) \qquad (60)\\
= \;& -2\sum_{t=R}^{T-1}(u_{t+1}x'_{1,t}(\hat{\beta}_{1,R} - \beta_1^*) - u_{t+1}x'_{2,t}(\hat{\beta}_{2,R} - \beta_2^*))\\
&+\sum_{t=R}^{T-1}((\hat{\beta}_{1,R} - \beta_1^*)'x_{1,t}x'_{1,t}(\hat{\beta}_{1,R} - \beta_1^*) - (\hat{\beta}_{2,R} - \beta_2^*)'x_{2,t}x'_{2,t}(\hat{\beta}_{2,R} - \beta_2^*))\\
= \;& 2(\frac{P}{R})^{1/2}(P^{-1/2}\sum_{t=R}^{T-1}u_{t+1}x'_{2,t})(-J(R^{-1}\sum_{s=1}^{R-1}x_{1,s}x'_{1,s})^{-1}J'\\
&+(R^{-1}\sum_{s=1}^{R-1}x_{2,s}x'_{2,s})^{-1})(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s})\\
&-(\frac{P}{R})(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s})(-J(R^{-1}\sum_{s=1}^{R-1}x_{1,s}x'_{1,s})^{-1}(P^{-1}\sum_{t=R}^{T-1}x_{1,t}x'_{1,t})(R^{-1}\sum_{s=1}^{R-1}x_{1,s}x'_{1,s})^{-1}J'\\
&+(R^{-1}\sum_{s=1}^{R-1}x_{2,s}x'_{2,s})^{-1}(P^{-1}\sum_{t=R}^{T-1}x_{1,t}x'_{1,t})(R^{-1}\sum_{s=1}^{R-1}x_{2,s}x'_{2,s})^{-1})(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s}).
\end{aligned}$$

So far we have used only algebra. In order to derive the asymptotic distribution of $\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2)$ we need to fall back on the assumptions in Clark and McCracken (2001) loosely presented in section 3.1.2, which for this simple case are closely related to those in West (1996): we need to assume that the sequence $(u_{s+1}, x_s)'$ is covariance stationary, mixing, and has bounded fourth moments. With these assumptions in hand it is clear that

$$\begin{aligned}
&\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) \qquad (61)\\
= \;& 2\pi^{1/2}(P^{-1/2}\sum_{t=R}^{T-1}u_{t+1}x'_{2,t})(-J(Ex_{1,s}x'_{1,s})^{-1}J' + (Ex_{2,s}x'_{2,s})^{-1})(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s})\\
&-\pi(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s})(-J(Ex_{1,s}x'_{1,s})^{-1}J' + (Ex_{2,s}x'_{2,s})^{-1})(R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s}) + o_p(1).
\end{aligned}$$

If we let both P and R tend to infinity, then $P^{-1/2}\sum_{t=R}^{T-1}u_{t+1}x_{2,t}$ and $R^{-1/2}\sum_{s=1}^{R-1}u_{s+1}x_{2,s}$ converge in distribution to $S_{hh}^{1/2}\tilde{W}_1$ and $S_{hh}^{1/2}\tilde{W}_2$, respectively, where $\tilde{W}_1$ and $\tilde{W}_2$ denote $(k \times 1)$ independent standard normal variates and S_{hh} denotes their (common) asymptotic variance. We therefore conclude that the asymptotic distribution of $\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2)$

takes the form

$$\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) \quad \rightarrow^d \quad 2\pi^{1/2}\tilde{W}_1' S_{hh}^{1/2}(-J(Ex_{1,s}x_{1,s}')^{-1}J' + (Ex_{2,s}x_{2,s}')^{-1})S_{hh}^{1/2}\tilde{W}_2$$
$$-\pi\tilde{W}_2' S_{hh}^{1/2}(-J(Ex_{1,s}x_{1,s}')^{-1}J' + (Ex_{2,s}x_{2,s}')^{-1})S_{hh}^{1/2}\tilde{W}_2.$$

Note that, under the recursive and rolling estimation schemes, the test statistics consist of partial sums that make the asymptotic distributions functions of Brownian motion instead of normal variates.

The above distribution is non-standard and involves the application-dependent (unknown, although estimable) parameters $Ex_{2,s}x_{2,s}'$ and S_{hh}. For that reason Clark and McCracken (20011b) recommend the bootstrap laid out in section 3.1.4 when conducting inference. However, in the special case in which the model errors u_{t+1} are conditionally homoskedastic and serially uncorrelated, a slightly modified version of this statistic has an asymptotic distribution that simplifies considerably, such that

$$(\sum_{t=R}^{T-1}\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2)/\hat{\sigma}_2^2 \rightarrow^d \qquad 2\Gamma_1 - \Gamma_2 \tag{62}$$
$$= \quad 2\pi^{1/2}W_1'W_2 - \pi W_2'W_2 \tag{63}$$

where W_i, $i = 1, 2$, denote $(k_w \times 1)$ independent standard normal vectors. While this distribution remains non-standard, it is free of nuisance parameters and can be readily simulated for a given value of π and dimension of x_w. The fact that this distribution does not involve stochastic integrals (as discussed in section 3.1.2) is a by-product of having used the fixed scheme to estimate model parameters. Were we to use the recursive scheme we would obtain the results presented in equation (5) of section 3.1.2.

11.3 Test of zero mean prediction error: Giacomini and White (2006)

Consider again the test of zero mean prediction error described in the previous section but now suppose that the parameter estimates used to construct the forecasts come from the rolling scheme and hence $R^{1/2}(\hat{\beta}_t - \beta^*) = (R^{-1}\sum_{s=t-R+1}^{t-1} x_s x_s')^{-1}(R^{-1/2}\sum_{s=t-R+1}^{t-1} u_{s+1}x_s)$. Straightforward algebra give us

$$P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1} = P^{-1/2}\sum_{t=R}^{T-1}y_{t+1} - x_t'\hat{\beta}_t = P^{-1/2}\sum_{t=R}^{T-1}(y_{t+1} - x_t'\beta^*) - x_t'(\hat{\beta}_t - \beta^*)$$
$$= P^{-1/2}\sum_{t=R}^{T-1}(u_{t+1} - x_t'(\hat{\beta}_t - \beta^*)).$$

So far this is just algebra. For this to be asymptotically normal we need to refer back to the assumptions made in Giacomini and White (2006). First recall that R is finite regardless of the overall sample size T, whereas P is assumed to diverge to infinity. This is crucial to their asymptotics because it implies that we can treat the sequence $\hat{\beta}_t - \beta^*$ as just another sequence of random variables without the added property that it is converging to zero. A central limit theorem can then be applied directly to $P^{-1/2}\sum_{t=R}^{T-1}(u_{t+1} - x_t'(\hat{\beta}_t - \beta^*))$ if we are willing to assume that the sequence $u_{t+1} - x_t'(\hat{\beta}_t - \beta^*)$ (on average) has a zero mean, and satisfies mild mixing and moment conditions. With these assumptions in hand we have

$$P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1} \rightarrow^d N(0, S_{\hat{f}\hat{f}}), \tag{64}$$

where $S_{\hat{f}\hat{f}} = \lim Var(P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1})$. Note that this is not the same asymptotic distribution as that given in equations (55) and (56) above. The difference arises due to the difference in the two null hypotheses as well as the difference in the type of assumptions being made on the data. The results in equations (55) and (56) are based on the null hypothesis $H_0 : Eu_{t+1} = 0$ for all t. The "all t" part is *imposed* by the additional assumptions that the observables are covariance stationary and the model includes an intercept. In contrast, the null hypothesis under the Giacomini and White framework is actually $\lim_{P\rightarrow\infty} E(P^{-1/2}\sum_{t=R}^{T-1}\hat{u}_{t+1}) = 0$, which is a much less stringent hypothesis. Note that Giacomini and White do not assume that the observables are covariance stationary — only that they are $I(0)$. Hence it might be that the population-level model errors u_{t+1} are zero mean but there is no requirement that is the case for the asymptotics to hold.

References

Amato, Jeffery D., and Norman R. Swanson (2001) "The Real Time Predictive Content of Money for Output," *Journal of Monetary Economics*, 48, 3-24.

Anatolyev, Stanislav (2007), "Inference About Predictive Ability When There Are Many Predictors," manuscript, New Economic School

Andrews, Donald W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica* 59, 1465-1471.

Andrews, Donald W.K., and J. Christopher Monahan (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica* 60, 953-966.

Aruoba, S. Boragan (2008), "Data Revisions Are Not Well-Behaved," *Journal of Money, Credit, and Banking* 40, 319-341.

Ashley, Richard, Clive W.J. Granger, and Richard L. Schmalensee (1980), "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica* 48, 1149-1167.

Busetti, Fabio, Juri Marcucci, and Giovanni Veronese (2009), "Comparing Forecast Accuracy: A Monte Carlo Investigation," Bank of Italy Working paper no. 723.

Calhoun, Gray (2011), "Out-of-Sample Comparisons of Overfit Models," manuscript, Iowa State University.

Chatfield C. (1995), "Model uncertainty, data mining and statistical inference," *Journal of the Royal Statistical Society, Series A*, 158, 419-466.

Chen, Yi-Ting (2011), "Moment Tests for Density Forecast Evaluation in the Presence of Parameter Estimation Uncertainty," *Journal of Forecasting*, 30, 409-450.

Chen, Yu-Chen, Kenneth S. Rogoff, and Barbara Rossi (2010), "Can Exchange Rates Forecast Commodity Prices?" *Quarterly Journal of Economics*, 125, 1145-1194.

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

Christoffersen, Peter, Eric Ghysels, and Norman R. Swanson (2002), "Let's Get 'Real' About Using Economic Data," *Journal of Empirical Finance*, 9, 343-360.

Clark, Todd E. (1999), "Finite-Sample Properties of Tests for Equal Forecast Accuracy," *Journal of Forecasting* 18, 489-504.

Clark, Todd E. (2004), "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting* 23, 115-139.

Clark, Todd E., and Taeyoung Doh (2011), "A Bayesian Evaluation of Alternative Models of Trend Inflation," manuscript, Federal Reserve Bank of Cleveland.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2005a), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24, 369-404.

Clark, Todd E., and Michael W. McCracken (2005b), "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics* 124, 1-31.

Clark, Todd E., and Michael W. McCracken (2009), "Tests of Equal Predictive Ability with Real-Time Data," *Journal of Business and Economic Statistics* 27, 441-454.

Clark, Todd E., and Michael W. McCracken (2011a), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," manuscript, Federal Reserve Bank of St. Louis, January.

Clark, Todd E., and Michael W. McCracken (2011b), "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business and Economic Statistics*, forthcoming.

Clark, Todd E., and Michael W. McCracken (2011c), "Testing for Unconditional Predictive Ability," in *Oxford Handbook of Economic Forecasting*, Michael P. Clements and David F. Hendry, eds., Oxford: Oxford University Press, forthcoming.

Clark, Todd E., and Michael W. McCracken (2012), "Tests of Equal Forecast Accuracy for Overlapping Models," manuscript, Federal Reserve Bank of St. Louis.

Clark, Todd E., and Kenneth D. West (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* 135, 155-186.

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

Clements, Michael P., and Ana Beatriz Galvao, "Real-time Forecasting of Inflation and Output Growth in the Presence of Data Revisions," manuscript,

Corradi, Valentina, and Walter Distaso (2011), "Multiple Forecast Model Evaluation," in *Oxford Handbook of Economic Forecasting*, Michael P. Clements and David F. Hendry, eds., Oxford: Oxford University Press, forthcoming.

Corradi, Valentina, and Norman R. Swanson (2006), "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, C.W.J. Granger, G. Elliott and A. Timmermann, eds., Elsevier: Amsterdam, 197-284.

Corradi, Valentina, and Norman R. Swanson (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review* 48, 67-109.

Corradi, Valentina, and Norman R. Swanson (2012), "A Survey of Recent Advances in Forecast Accuracy Testing, with an Extension to Stochastic Dominance," forthcoming in *Recent Advances and Future Directions in Causality, Prediction and Specification Analysis: Essays in Honor of Halbert L. White, Jr.* .

Croushore, Dean (2006). Forecasting with Real–Time Macroeconomic Data. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting* (pp. 961-82). Amsterdam The Netherlands: North-Holland.

Croushore, Dean, and Tom Stark (2003), "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *The Review of Economics and Statistics*, 85, 605-617.

Davidson, Russell (1994), *Stochastic Limit Theory*, New York: Oxford University Press.

de Jong, Robert M., and James Davidson (2000), "The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes," *Econometric Theory* 16, 621-642.

Denton, Frank T. (1985), "Data Mining as an Industry," *Review of Economics and Statistics* 67, 124-127.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Diebold, Francis X., and Glenn D. Rudebusch (1991), "Forecasting Output with the Composite Leading Index: A Real-Time Analysis," *Journal of the American Statistical Association*, 86, 603-610.

Fair, Ray C., and Robert J. Shiller (1989), "The Informational Content of Ex Ante Forecasts," *Review of Economics and Statistics* 71, 325-331.

Fair, Ray C., and Robert J. Shiller (1990), "Comparing Information in Forecasts from Econometric Models," *American Economic Review* 80, 375-389.

Giacomini, Rafaella, and Barbara Rossi (2009), "Detecting and Predicting Forecast Breakdown," *Review of Economic Studies* 76, 669-705.

Giacomini, Rafaella, and Halbert White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Goyal, Amit, and Ivo Welch (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies* 21, 1455-1508.

Granziera, Eleonora, Kirstin Hubrich, and Roger H. Moon (2011), "A Predictability test for a Small Number of Nested Models," manuscript.

Hallman, Jeffrey J., Richard D. Porter, and David H. Small (1991), "Is the Price Level Tied to the M2 Monetary Aggregate in the Long Run?" *American Economic Review* 81, 841-858.

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

Hansen, Lars Peter (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029-1054.

Hansen, Peter Reinhard (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23, 365-380.

Hansen, Peter Reinhard, and Allan Timmermann (2011), "Choice of Sample Split in Out-of-Sample Forecast Evaluation," manuscript, Stanford University.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1997), "Testing the Equality

of Prediction Mean Squared Errors," *International Journal of Forecasting* 13, 281-91.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16, 254-259.

Hendry, David F. (2004), "Unpredictability and the Foundations of Economic Forecasting," manuscript, Nuffield College.

Hodrick, Robert J. (1992), "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," *Review of Financial Studies* 5, 357-386.

Hoogerheide, Lennart, Francesco Ravazzolo, and Herman K. van Dijk (2012), "Comment on 'Forecast Rationality Tests Based on Multi-Horizon Bounds'," *Journal of Business and Economic Statistics*, 30, 30-33.

Hoover, Kevin D., and S.J. Perez (1999), "Data mining reconsidered: encompassing and the general–to–specific approach to specification search," *Econometrics Journal* 2, 167-191.

Howrey, E. Philip (1978), "The Use of Preliminary Data in Econometric Forecasting," *Review of Economics and Statistics*, 60, 193-200.

Hubrich, Kirstin, and Kenneth D. West (2010), "Forecast Evaluation of Small Nested Model Sets," *Journal of Applied Econometrics* 25, 574-594.

Inoue, Atsushi, and Lutz Kilian (2004), "In–Sample or Out–of–Sample Tests of Predictability? Which One Should We Use?" *Econometric Reviews* 23, 371-402.

Inoue, Atsushi, and Barbara Rossi (2005), "Recursive Predictability Tests for Real-Time Data," *Journal of Business and Economic Statistics*, 23, 336-345.

Kilian, Lutz (1998), "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics* 80, 218-230.

Kilian, Lutz (1999), "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?" *Journal of Applied Econometrics* 14, 491-510.

Kishor, N. Kundan, and Evan F. Koenig (2011), "VAR Estimation and Forecasting When Data Are Subject to Revision," *Journal of Business and Economic Statistics*, forthcoming.

Koenig, Evan F., Shelia Dolmas, and Jeremy Piger (2003), "The Use and Abuse of Real-Time Data in Economic Forecasting," *The Review of Economics and Statistics*, 85, 618-628.

Kozicki, Sharon, and Peter A. Tinsley (2001), "Shifting Endpoints in the Term Structure of Interest Rates," *Journal of Monetary Economics* 47, 613-652.

Leamer, Edward E. (1978), *Specification Searches: Ad Hoc Inference with Experimental Data*, Wiley: New York.

Lovell, Michael C. (1983), "Data Mining," *Review of Economics and Statistics* 65, 1-12.

Lo, Andrew W., and A. Craig MacKinlay (1990), "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies* 3, 431-467.

Mankiw, N. Gregory, David E. Runkle, and Matthew D. Shapiro (1984), "Are Preliminary Announcements of the Money Stock Rational Forecasts?" *Journal of Monetary Economics*, 14, 15-27.

Mariano, Roberto S., and Daniel Preve (2012), "Model-Free Tests for Multiple Forecast Comparison," *Journal of Econometrics*, 169, 123-130.

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

McCracken, Michael W. (2004), "Parameter Estimation Error and Tests of Equal Forecast Accuracy Between Non-nested Models," *The International Journal of Forecasting*, 20, 503-514.

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Meese, Richard, and Kenneth S. Rogoff (1983), "Empirical exchange rate models of the seventies: Do they fit out of sample?" *Journal of International Economics* 14, 3-24.

Meese, Richard, and Kenneth S. Rogoff (1988), "Was it Real? The Exchange Rate–Interest Differential Relation Over the Modern Floating-Rate Period," *Journal of Finance* 43, 933-948.

Newey, Whitney K., and Kenneth D. West (1987), "A simple, positive semi–definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55:703–708.

Newey, Whitney K., and Kenneth D. West (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631-53.

Orphanides, Athanasios, and Simon van Norden (2005), "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," *Journal of Money, Credit, and Banking*, 37, 583-601.

Patton, Andrew J., and Allan Timmermann (2012), "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business and Economic Statistics*, 30, 1-17.

Rapach, David, and Mark E. Wohar (2006), "In-Sample vs. Out-of-Sample Tests of Stock Return Predictability in the Context of Data Mining," *Journal of Empirical Finance* 13, 231-247.

Rossi, Barbara, and Atsushi Inoue (2011), "Out-of-Sample Forecast Tests Robust to the Window Size Choice," manuscript, Duke University.

Song, Kyungchul (2012), "Testing Predictive Ability and Power Robustification," textitJournal of Business and Economic Statistics, 30, 288-296.

Stambaugh, Robert F. (1999), "Predictive Regressions," *Journal of Financial Economics* 54, 375-421.

Stock, James H., and Mark W. Watson (1996), "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14, 11-30.

Stock, James H., and Mark W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41, 788-829.

Stock, James H., and Mark W. Watson (2007), "Has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit, and Banking* 39, 3-33.

Vuong, Quang H., (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica* 57, 307-333.

West, Kenneth D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

West, Kenneth D. (1997), "Another Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Journal of Econometrics* 76, 171-191.

West, Kenneth D. (2006), "Forecast Evaluation," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., Timmermann, A. (eds), North Holland.

West, Kenneth D. (2012), "Comment on 'Forecast Rationality Tests Based on Multi-Horizon Bounds'," *Journal of Business and Economic Statistics*, 30, 34-35.

West, Kenneth D., and Michael W. McCracken (1998), "Regression-based Tests of Predictive Ability," *International Economic Review* 39, 817-840.

White, Halbert (2000), "A Reality Check For Data Snooping," *Econometrica* 68, 1097-1127.

Wilson, Edwin B. (1934), "The Periodogram of American Business Activity," *The Quarterly Journal of Economics,* 48, 375–417.

Wright, Jonathan H. (1997), "The Limiting Distribution of Post-sample Stability Tests for GMM Estimation when the Potential Break Date is Unknown," *Oxford Bulletin of Economics and Statistics,* 59, 299–303.

Table 3: Tests of Equal Forecast Accuracy, Non-Nested Models of Inflation, Current Vintage Data

	Recursive scheme	*Rolling scheme*
Horizon = 1Q		
RMSE, model with GDP growth	0.809	0.815
RMSE, model with GDP gap	0.831	0.836
MSE-t test (p-value)	-1.125 (0.260)	-1.458 (0.145)
GW test for conditional EPA (p-value)	1.603 (0.449)	2.106 (0.349)
Horizon = 4Q		
RMSE, model with GDP growth	0.616	0.665
RMSE, model with GDP gap	0.873	0.812
MSE-t test (p-value)	-1.428 (0.153)	-1.692 (0.091)
GW test for conditional EPA (p-value)	2.952 (0.229)	3.087 (0.214)

Notes:
1. As described in section 3.3, forecasts of inflation (defined as 400 times the log difference of the GDP price index) are generated from models of the form of equation (15). The model includes two lags of inflation at the one-quarter forecast horizon and one lag of inflation at the four-quarter horizon. One model includes GDP growth; the other includes the GDP gap. The forecast sample is 1985:Q1 + horizon − 1 through 2011:Q4.
2. The test statistic MSE-t is defined in Table 1. The GW test for conditional EPA is defined in section 4.1. The variance estimates needed for the test statistics are computed with a rectangular kernel and bandwidth of horizon less one; the variance used in the MSE-t test also includes the finite-sample adjustment of Harvey, Leybourne, and Newbold (1997). The p-values of the MSE-t and GW tests are obtained from, respectively, the normal and χ^2 distributions.

Table 4: Tests of Equal Forecast Accuracy, Nested Models of Inflation, Current Vintage Data

	Recursive		*Rolling*	
	Model with GDP growth	Model with GDP gap	Model with GDP growth	Model with GDP gap
Horizon = 1Q				
RMSE/RMSE of AR model	0.980	1.007	0.980	1.005
MSE-t	0.898	-0.170	0.807	-0.215
p-value, Normal	0.185	0.568	0.210	0.585
p-value, FRBS, population EPA	0.072	0.248	0.049	0.161
p-value, FRBS, finite-sample EPA	0.213	0.443	0.177	0.346
MSE-F	4.463	-1.398	4.508	-1.058
MSE-F p-value, FRBS, population EPA	0.026	0.579	0.020	0.215
MSE-F p-value, FRBS, finite-sample EPA	0.137	0.663	0.126	0.387
Horizon = 4Q				
RMSE/RMSE of AR model	0.920	1.304	0.923	1.127
MSE-t	0.564	-0.906	0.748	-0.734
p-value, Normal	0.287	0.818	0.227	0.769
p-value, FRBS, population EPA	0.201	0.455	0.153	0.314
p-value, FRBS, finite-sample EPA	0.385	0.830	0.321	0.791
MSE-F	19.017	-43.240	18.191	-22.301
p-value, FRBS, population EPA	0.039	0.987	0.040	0.916
p-value, FRBS, finite-sample EPA	0.176	0.993	0.197	0.948

Notes:
1. As described in section 3.3, forecasts of inflation (defined as 400 times the log difference of the GDP price index) are generated from models of the form of equations (14) and (15), where equation (14) is the benchmark and equation (15) is the alternative. The models include two lags of inflation at the one-quarter forecast horizon and one lag of inflation at the four-quarter horizon. One alternative model includes GDP growth; the other alternative model includes the GDP gap. Each is separately tested against the AR model benchmark. The forecast sample is 1985:Q1 + horizon − 1 through 2011:Q4.
2. The MSE-t and MSE-F test statistics are defined in Table 1. The variance estimates needed for the MSE-t test statistics are computed with a rectangular kernel and bandwidth of horizon less one and include the finite-sample adjustment of Harvey, Leybourne, and Newbold (1997). Because the models are nested, all of the tests are one-sided, rejecting the null only if the alternative model is more accurate.
3. The table reports p-values obtained several different ways. Under the null of equal accuracy in population, the table provides p-values computed under the fixed regressor bootstrap described in section 3.1.4. Under the null of equal accuracy in the finite sample, the table provides p-values for the MSE-t test compared against standard normal critical values (valid for the rolling estimation scheme based on the asymptotics of Giacomini and White (2006)) and p-values for both equal MSE tests based on the finite-sample fixed regressor bootstrap described in section 3.2.2. The number of bootstrap draws is 4999.
4. The RMSEs of the benchmark AR model are as follows: one-quarter horizon, 0.826 for the recursive scheme and 0.832 for the rolling scheme; four-quarter horizon, 0.670 for the recursive scheme and 0.720 for the rolling scheme.

Table 5: Tests of Equal Forecast Accuracy, Overlapping Models of Inflation, Current Vintage Data

	Horizon = 1Q	*Horizon = 4Q*
RMSE, model with GDP growth	0.809	0.616
RMSE, model with GDP gap	0.831	0.873
$\hat{S}_{dd}$ (p-value)	0.110 (0.026)	7.112 (0.000)
MSE-t	-1.125	-1.428
90% bootstrap critical values for MSE-t	-1.971, 1.305	-1.881, 1.310
95% bootstrap critical values for MSE-t	-2.279, 1.563	-2.202, 1.571
1-step procedure 90% critical values for MSE-t	-1.971, 1.645	-1.881, 1.645
1-step procedure 95% critical values for MSE-t	-2.279, 1.960	-2.202, 1.960

Notes:
1. As described in section 3.3, forecasts of inflation (defined as 400 times the log difference of the GDP price index) are generated from models of the form of equation (15). The model includes two lags of inflation at the one-quarter forecast horizon and one lag of inflation at the four-quarter horizon. One model includes GDP growth; the other includes the GDP gap. The forecast sample is 1985:Q1 + horizon − 1 through 2011:Q4.
2. The test statistic MSE-t is defined in Table 1. The $\hat{S}_{dd}$ statistic is defined in section 3.1.3. The variance estimates needed for the test statistics are computed with a rectangular kernel and bandwidth of horizon less one; the variance used in the MSE-t test also includes the finite-sample adjustment of Harvey, Leybourne, and Newbold (1997).
3. The fixed regressor bootstrap and the one-step testing procedure are described in section 3.1.3. The number of bootstrap draws is 4999.

Table 6: Tests of Equal Forecast Accuracy,
Multiple Nested Models of Inflation, Current Vintage Data

Model *predictors*	*RMSE/* *RMSE AR*	MSE-F (p-value)	MSE-t (p-value)
Horizon = 1Q			
Best model (reality check)	0.980	4.463 (0.063)	0.898 (0.152)
GDP growth	0.980	4.463 (0.025)	0.898 (0.084)
GDP gap	1.007	-1.398 (0.544)	-0.170 (0.223)
Capacity utilization	1.024	-4.960 (0.919)	-0.659 (0.411)
GDP growth, F&E inflation	1.024	-5.017 (0.842)	-0.631 (0.446)
GDP gap, F&E inflation	1.038	-7.781 (0.920)	-0.743 (0.427)
Capacity utilization, F&E inflation	1.046	-9.238 (0.951)	-1.074 (0.565)
GDP growth, GDP gap, F&E inflation	1.054	-10.768 (0.951)	-1.053 (0.530)
Horizon = 4Q			
Best model (reality check)	0.920	19.017 (0.076)	0.564 (0.392)
GDP growth	0.920	19.017 (0.033)	0.564 (0.224)
GDP gap	1.304	-43.240 (0.997)	-0.906 (0.537)
Capacity utilization	1.407	-51.944 (1.000)	-1.548 (0.871)
GDP growth, F&E inflation	1.173	-28.739 (0.955)	-1.225 (0.683)
GDP gap, F&E inflation	1.440	-54.351 (0.997)	-1.379 (0.683)
Capacity utilization, F&E inflation	1.446	-54.799 (0.998)	-2.086 (0.905)
GDP growth, GDP gap, F&E inflation	1.448	-54.899 (0.998)	-1.380 (0.672)

Notes:
1. As described in section 5.4, forecasts of inflation (defined as 400 times the log difference of the GDP price index) are generated from models of the form of equations (14) and (15), where equation (14) is the benchmark and equation (15), modified to include some additional predictors, is the alternative. The models include two lags of inflation at the one-quarter forecast horizon and one lag of inflation at the four-quarter horizon. The seven different alternative models considered include the predictors listed in the first column of the table.
2. This table provides pairwise tests and reality check (best model) tests of equal forecast accuracy. For each alternative model, the table reports the ratio of the alternative model's RMSE to the null model's forecast RMSE and bootstrapped p-values for the null hypothesis of equal accuracy, for the MSE-t and MSE-F test statistics, defined in Table 1. The variance estimates needed for the MSE-t test statistics are computed with a rectangular kernel and bandwidth of horizon less one and include the finite-sample adjustment of Harvey, Leybourne, and Newbold (1997). Because the models are nested, all of the tests are one-sided, rejecting the null only if the alternative model is more accurate. The top row of each panel gives the test statistics of the best models and reality check p-values. Sections 3.1.4 and 5.4 describe the bootstrap. The number of bootstrap draws is 4999.
3. The RMSEs of the benchmark AR model are 0.826 at the one-quarter horizon and 0.670 at the four-quarter horizon.

Table 7: Tests of Equal Forecast Accuracy,
Non-Nested and Nested Models of Inflation, Real-Time Data

sample	MSE_1	MSE_2	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	MSE-$t(S_{dd})$	MSE-$t(\Omega)$	MSE-F
Non-nested models: GDP gap (model 1) vs. GDP growth (model 2)							
actual inflation$_t$ = estimate published in $t+2$							
Horizon = 1Q	1.139	1.184	0.031	0.036	-1.472	-1.254	NA
Horizon = 4Q	0.427	0.356	0.071	0.079	0.999	0.895	NA
actual inflation$_t$ = estimate published in $t+5$							
Horizon = 1Q	1.094	1.123	0.035	0.039	-0.829	-0.735	NA
Horizon = 4Q	0.480	0.398	0.078	0.088	1.050	0.934	NA
actual inflation$_t$ = estimate published in $t+13$							
Horizon = 1Q	1.047	1.046	0.034	0.039	0.028	0.024	NA
Horizon = 4Q	0.576	0.450	0.103	0.114	1.222	1.105	NA
Nested models: AR (model 1) vs. GDP gap (model 2)							
actual inflation$_t$ = estimate published in $t+2$							
Horizon = 1Q	1.223	1.139	0.040	0.008	2.088^c	10.271^c	7.007^c
Horizon = 4Q	0.518	0.427	0.090	0.039	1.017^b	2.347^c	19.856^b
actual inflation$_t$ = estimate published in $t+5$							
Horizon = 1Q	1.163	1.094	0.044	0.005	1.559^c	14.281^c	6.015^c
Horizon = 4Q	0.544	0.480	0.098	0.053	0.657^a	1.219	12.439^b
actual inflation$_t$ = estimate published in $t+13$							
Horizon = 1Q	1.072	1.047	0.046	0.012	0.549^a	2.062^c	2.317^b
Horizon = 4Q	0.545	0.576	0.121	0.085	-0.255	-0.367	-5.007
Nested models: AR (model 1) vs. GDP growth (model 2)							
actual inflation$_t$ = estimate published in $t+2$							
Horizon = 1Q	1.223	1.184	0.025	0.014	1.534^a	2.725^c	3.093^b
Horizon = 4Q	0.518	0.356	0.086	0.008	1.895^b	21.617^c	42.426^c
actual inflation$_t$ = estimate published in $t+5$							
Horizon = 1Q	1.163	1.123	0.022	0.014	1.781^b	2.870^c	3.384^b
Horizon = 4Q	0.544	0.398	0.077	0.003	1.892^b	44.069^c	34.233^c
actual inflation$_t$ = estimate published in $t+13$							
Horizon = 1Q	1.072	1.046	0.026	0.007	1.001	3.823^c	2.405^b
Horizon = 4Q	0.545	0.450	0.081	0.013	1.178	7.331^c	19.725^c

Notes:
1. As described in section 6.3, real time forecasts of inflation in the GDP price index are generated from models of the form of equations (14) and (15). The forecasts in the non-nested comparison are generated from equation (15), with model 1 using x_t = the output gap and model 2 using x_t = four-quarter GDP growth. The forecasts in the nested comparison are generated from equations (14) (model 1) and (15) (model 2). The models include two lags of inflation at the one-quarter forecast horizon and one lag of inflation at the four-quarter horizon. The models are estimated recursively.
2. The MSEs are based on forecasts computed with various definitions of actual inflation used in computing forecast errors. The first panel takes actual to be the second available estimate of inflation; the next the fifth available estimate; and so on.
3. The columns MSE-$t(S_{dd})$ and MSE-$t(\Omega)$ report t-statistics for the difference in MSEs computed with the variances $\hat{S}_{dd}$ and $\hat{\Omega}$, respectively. In the non-nested comparison, the variance Ω is defined as $\hat{S}_{dd} + 2\hat{\Pi}(\hat{F}\hat{B}\hat{S}_{dh} + \hat{F}\hat{B}\hat{S}_{hh}\hat{B}\hat{F}')$. The non-nested tests are compared against standard normal critical values. In the nested comparison, $\Omega = 2\hat{\Pi}\hat{F}(-J\hat{B}_1J' + \hat{B}_2)\hat{S}_{hh}(-J\hat{B}_1J' + \hat{B}_2)\hat{F}'$. In the nested model comparisons, MSE-$t(S_{dd})$ and MSE-F are compared against critical values simulated as in Clark and McCracken (2005a), and the MSE-$t(\Omega)$ statistic is compared against standard normal critical values. Test statistics rejecting the null of equal accuracy at significance levels of 10%, 5%, and 1% are denoted by superscripts of, respectively, ^{a}, ^{b}, and ^{c}.

Table 9: Monte Carlo Results on Size, DGP 1: Equal Accuracy in Population

(nominal size = 10%)

statistic	*HAC estimator*	*source of critical values*	R=40 $\tilde{P}$=80	R=40 $\tilde{P}$=120	R=80 $\tilde{P}$=20	R=80 $\tilde{P}$=40	R=80 $\tilde{P}$=80	R=80 $\tilde{P}$=120	R=120 $\tilde{P}$=40	R=120 $\tilde{P}$=80
horizon = 4										
MSE-F	NA	FRBS: no pred.	0.105	0.104	0.103	0.106	0.108	0.108	0.103	0.108
MSE-t	NW	FRBS: no pred.	0.099	0.102	0.101	0.103	0.102	0.104	0.103	0.101
MSE-t	NW	normal	0.025	0.019	0.131	0.077	0.042	0.029	0.087	0.047
MSE-t	rectangular	normal	0.022	0.015	0.120	0.067	0.038	0.025	0.079	0.042
MSE-t	HLN	normal	0.020	0.013	0.088	0.055	0.033	0.023	0.065	0.036
MSE-t	West	normal	0.021	0.014	0.144	0.066	0.033	0.024	0.076	0.038
MSE-t	QS	normal	0.016	0.011	0.082	0.047	0.026	0.018	0.057	0.028
CW-t	NW	FRBS: no pred.	0.094	0.102	0.099	0.096	0.099	0.103	0.103	0.102
CW-t	NW	normal	0.096	0.093	0.188	0.136	0.106	0.097	0.139	0.111
CW-t	rectangular	normal	0.085	0.080	0.173	0.124	0.094	0.086	0.129	0.098
CW-t	HLN	normal	0.078	0.078	0.129	0.104	0.088	0.082	0.111	0.091
CW-t	West	normal	0.080	0.080	0.196	0.120	0.088	0.082	0.121	0.092
CW-t	QS	normal	0.066	0.067	0.121	0.089	0.071	0.068	0.094	0.071
horizon = 8										
MSE-F	NA	FRBS: no pred.	0.110	0.104	0.106	0.109	0.111	0.100	0.107	0.102
MSE-t	NW	FRBS: no pred.	0.112	0.098	0.100	0.112	0.108	0.095	0.108	0.098
MSE-t	NW	normal	0.048	0.025	0.189	0.117	0.060	0.033	0.125	0.067
MSE-t	rectangular	normal	0.049	0.028	0.147	0.113	0.061	0.032	0.119	0.067
MSE-t	HLN	normal	0.040	0.024	0.081	0.083	0.048	0.027	0.086	0.057
MSE-t	West	normal	0.057	0.026	0.254	0.151	0.068	0.033	0.157	0.073
MSE-t	QS	normal	0.027	0.015	0.098	0.063	0.036	0.023	0.072	0.043
CW-t	NW	FRBS: no pred.	0.103	0.097	0.105	0.103	0.106	0.095	0.103	0.093
CW-t	NW	normal	0.127	0.104	0.254	0.184	0.136	0.104	0.183	0.125
CW-t	rectangular	normal	0.127	0.101	0.199	0.179	0.133	0.104	0.172	0.122
CW-t	HLN	normal	0.109	0.091	0.117	0.136	0.115	0.091	0.131	0.104
CW-t	West	normal	0.133	0.101	0.320	0.217	0.140	0.102	0.216	0.129
CW-t	QS	normal	0.084	0.071	0.140	0.110	0.089	0.072	0.115	0.082

Notes:
1. The data generating process is defined in equation (29). In these experiments, the coefficients $b_{ij} = 0$ for all i, j, such that the competing forecasting models are equally accurate in population, but not the finite sample.
2. For each artificial data set, forecasts of $y_{t+\tau}$ (where τ denotes the forecast horizon) are formed recursively using estimates of equations (30) and (31). These forecasts are then used to form the indicated test statistics, defined in Table 1, using the indicated HAC estimator, defined in Section 7.1.3. R and $\tilde{P}$ refer to the number of in–sample observations and τ-step ahead forecasts, respectively (where $\tilde{P} = P + \tau - 1$, and P denotes the sample size used in the paper's theory).
3. In each Monte Carlo replication, the simulated test statistics are compared against standard normal critical values and critical values bootstrapped using the no-predictability fixed regressor bootstrap, using a significance level of 10%. Section 3.1.4 describes the bootstrap procedure.
4. The number of Monte Carlo simulations is 5000; the number of bootstrap draws is 499.

Table 10: Monte Carlo Results on Size, DGP 2: Equal Accuracy in Population

(nominal size = 10%)

statistic	*HAC estimator*	*source of critical values*	R=40 $\tilde{P}$=80	R=40 $\tilde{P}$=120	R=80 $\tilde{P}$=20	R=80 $\tilde{P}$=40	R=80 $\tilde{P}$=80	R=80 $\tilde{P}$=120	R=120 $\tilde{P}$=40	R=120 $\tilde{P}$=80
horizon = 4										
MSE-F	NA	FRBS: no pred.	0.108	0.097	0.111	0.104	0.102	0.107	0.102	0.107
MSE-t	NW	FRBS: no pred.	0.106	0.100	0.117	0.101	0.098	0.108	0.104	0.102
MSE-t	NW	normal	0.011	0.005	0.113	0.049	0.022	0.014	0.065	0.030
MSE-t	rectangular	normal	0.009	0.004	0.100	0.045	0.017	0.010	0.057	0.024
MSE-t	HLN	normal	0.007	0.004	0.071	0.035	0.014	0.009	0.045	0.021
MSE-t	West	normal	0.008	0.004	0.109	0.043	0.015	0.009	0.055	0.020
MSE-t	QS	normal	0.006	0.003	0.073	0.032	0.010	0.008	0.036	0.017
CW-t	NW	FRBS: no pred.	0.096	0.090	0.110	0.094	0.092	0.098	0.097	0.097
CW-t	NW	normal	0.119	0.101	0.205	0.147	0.116	0.112	0.149	0.120
CW-t	rectangular	normal	0.103	0.086	0.189	0.132	0.097	0.097	0.129	0.104
CW-t	HLN	normal	0.094	0.079	0.146	0.111	0.088	0.092	0.112	0.093
CW-t	West	normal	0.098	0.085	0.198	0.121	0.092	0.091	0.124	0.096
CW-t	QS	normal	0.074	0.065	0.140	0.092	0.072	0.073	0.092	0.075
horizon = 8										
MSE-F	NA	FRBS: no pred.	0.111	0.111	0.118	0.117	0.114	0.100	0.120	0.114
MSE-t	NW	FRBS: no pred.	0.107	0.109	0.113	0.112	0.107	0.100	0.113	0.110
MSE-t	NW	normal	0.021	0.014	0.165	0.084	0.039	0.019	0.098	0.044
MSE-t	rectangular	normal	0.020	0.013	0.134	0.077	0.037	0.017	0.097	0.044
MSE-t	HLN	normal	0.017	0.011	0.073	0.053	0.030	0.014	0.069	0.034
MSE-t	West	normal	0.020	0.013	0.210	0.098	0.036	0.017	0.112	0.041
MSE-t	QS	normal	0.014	0.008	0.088	0.046	0.021	0.010	0.056	0.025
CW-t	NW	FRBS: no pred.	0.107	0.104	0.108	0.106	0.104	0.099	0.109	0.105
CW-t	NW	normal	0.150	0.130	0.277	0.194	0.155	0.126	0.202	0.152
CW-t	rectangular	normal	0.140	0.120	0.231	0.182	0.144	0.116	0.197	0.147
CW-t	HLN	normal	0.120	0.104	0.138	0.143	0.120	0.103	0.157	0.123
CW-t	West	normal	0.136	0.111	0.327	0.213	0.138	0.109	0.214	0.139
CW-t	QS	normal	0.100	0.086	0.159	0.124	0.095	0.082	0.122	0.097

Notes:
1. The data generating process is defined in equation (32), and the forecasting models are given in equations (33) and (34).
2. See the notes to Table 9.

Table 11: Monte Carlo Results on Size, DGP 1: Equal Accuracy in Finite Sample
(nominal size = 10%)

statistic	*HAC estimator*	*source of critical values*	R=40 $\tilde{P}$=80	R=40 $\tilde{P}$=120	R=80 $\tilde{P}$=20	R=80 $\tilde{P}$=40	R=80 $\tilde{P}$=80	R=80 $\tilde{P}$=120	R=120 $\tilde{P}$=40	R=120 $\tilde{P}$=80
horizon = 4										
MSE-F	NW	FRBS	0.147	0.143	0.131	0.136	0.126	0.131	0.116	0.123
MSE-F	rectangular	FRBS	0.135	0.128	0.121	0.127	0.115	0.119	0.107	0.111
MSE-F	West	FRBS	0.106	0.101	0.100	0.111	0.097	0.096	0.090	0.094
MSE-F	QS	FRBS	0.112	0.101	0.103	0.107	0.095	0.096	0.088	0.091
MSE-t	NW	FRBS	0.133	0.130	0.112	0.117	0.110	0.119	0.107	0.112
MSE-t	rectangular	FRBS	0.123	0.115	0.107	0.113	0.102	0.108	0.101	0.105
MSE-t	West	FRBS	0.098	0.089	0.099	0.100	0.087	0.091	0.094	0.092
MSE-t	QS	FRBS	0.104	0.092	0.097	0.102	0.086	0.087	0.092	0.090
MSE-t	NW	normal	0.119	0.094	0.204	0.157	0.107	0.096	0.146	0.115
MSE-t	rectangular	normal	0.105	0.077	0.186	0.143	0.094	0.085	0.134	0.104
MSE-t	HLN	normal	0.097	0.072	0.141	0.122	0.086	0.079	0.115	0.094
MSE-t	West	normal	0.099	0.074	0.201	0.138	0.087	0.081	0.130	0.097
MSE-t	QS	normal	0.077	0.057	0.139	0.107	0.072	0.064	0.097	0.079
horizon = 8										
MSE-F	NW	FRBS	0.153	0.149	0.131	0.147	0.143	0.132	0.123	0.129
MSE-F	rectangular	FRBS	0.151	0.145	0.128	0.141	0.141	0.124	0.118	0.122
MSE-F	West	FRBS	0.102	0.098	0.093	0.102	0.101	0.094	0.095	0.100
MSE-F	QS	FRBS	0.120	0.111	0.103	0.113	0.111	0.101	0.099	0.100
MSE-t	NW	FRBS	0.139	0.129	0.111	0.129	0.125	0.117	0.102	0.115
MSE-t	rectangular	FRBS	0.137	0.127	0.111	0.127	0.123	0.114	0.099	0.109
MSE-t	West	FRBS	0.096	0.089	0.092	0.103	0.097	0.092	0.084	0.094
MSE-t	QS	FRBS	0.109	0.097	0.097	0.109	0.104	0.097	0.087	0.097
MSE-t	NW	normal	0.149	0.116	0.270	0.219	0.143	0.117	0.182	0.144
MSE-t	rectangular	normal	0.146	0.109	0.216	0.199	0.142	0.113	0.171	0.141
MSE-t	HLN	normal	0.122	0.095	0.131	0.151	0.122	0.101	0.127	0.118
MSE-t	West	normal	0.147	0.108	0.322	0.240	0.141	0.110	0.207	0.145
MSE-t	QS	normal	0.101	0.080	0.158	0.133	0.096	0.080	0.112	0.095

Notes:
1. The data generating process is defined in equation (29). In these experiments, the coefficients b_{ij} are scaled such that the null and alternative models are expected to be equally accurate (on average) over the forecast sample.
2. For each artificial data set, forecasts of $y_{t+\tau}$ (where τ denotes the forecast horizon) are formed recursively using estimates of equations (30) and (31). These forecasts are then used to form the indicated test statistics, defined in Table 1, using the indicated HAC estimator, defined in Section 7.1.3. R and $\tilde{P}$ refer to the number of in–sample observations and τ-step ahead forecasts, respectively (where $\tilde{P} = P + \tau - 1$, and P denotes the sample size used in the paper's theory).
3. In each Monte Carlo replication, the simulated test statistics are compared against standard normal critical values and critical values bootstrapped using the no-predictability fixed regressor bootstrap, using a significance level of 10%. Section 3.1.4 describes the bootstrap procedure.
4. The number of Monte Carlo simulations is 5000; the number of bootstrap draws is 499.

Table 12: Monte Carlo Results on Size, DGP 2: Equal Accuracy in Finite Sample

(nominal size = 10%)

statistic	*HAC estimator*	*source of critical values*	R=40 $\tilde{P}$=80	R=40 $\tilde{P}$=120	R=80 $\tilde{P}$=20	R=80 $\tilde{P}$=40	R=80 $\tilde{P}$=80	R=80 $\tilde{P}$=120	R=120 $\tilde{P}$=40	R=120 $\tilde{P}$=80
horizon = 4										
MSE-F	NW	FRBS	0.191	0.185	0.133	0.154	0.160	0.174	0.135	0.154
MSE-F	rectangular	FRBS	0.165	0.160	0.118	0.137	0.142	0.151	0.118	0.135
MSE-F	West	FRBS	0.114	0.113	0.088	0.101	0.109	0.119	0.095	0.107
MSE-F	QS	FRBS	0.134	0.118	0.098	0.113	0.109	0.116	0.093	0.103
MSE-t	NW	FRBS	0.166	0.169	0.119	0.133	0.142	0.151	0.127	0.134
MSE-t	rectangular	FRBS	0.147	0.144	0.111	0.121	0.127	0.134	0.117	0.119
MSE-t	West	FRBS	0.100	0.105	0.096	0.100	0.105	0.105	0.104	0.099
MSE-t	QS	FRBS	0.121	0.111	0.101	0.106	0.107	0.107	0.102	0.096
MSE-t	NW	normal	0.125	0.111	0.214	0.163	0.132	0.120	0.169	0.136
MSE-t	rectangular	normal	0.108	0.091	0.187	0.146	0.115	0.102	0.153	0.115
MSE-t	HLN	normal	0.099	0.084	0.142	0.121	0.104	0.094	0.131	0.105
MSE-t	West	normal	0.106	0.086	0.197	0.137	0.105	0.095	0.141	0.106
MSE-t	QS	normal	0.075	0.062	0.139	0.101	0.081	0.071	0.107	0.079
horizon = 8										
MSE-F	NW	FRBS	0.174	0.170	0.142	0.154	0.158	0.147	0.133	0.135
MSE-F	rectangular	FRBS	0.162	0.154	0.137	0.145	0.146	0.135	0.122	0.125
MSE-F	West	FRBS	0.091	0.091	0.099	0.098	0.096	0.090	0.090	0.092
MSE-F	QS	FRBS	0.135	0.132	0.128	0.129	0.129	0.115	0.112	0.109
MSE-t	NW	FRBS	0.150	0.157	0.125	0.135	0.142	0.131	0.119	0.120
MSE-t	rectangular	FRBS	0.142	0.141	0.121	0.129	0.134	0.121	0.113	0.115
MSE-t	West	FRBS	0.089	0.088	0.099	0.099	0.101	0.089	0.091	0.090
MSE-t	QS	FRBS	0.122	0.123	0.115	0.121	0.120	0.106	0.104	0.101
MSE-t	NW	normal	0.151	0.130	0.290	0.213	0.157	0.124	0.207	0.145
MSE-t	rectangular	normal	0.140	0.118	0.239	0.195	0.145	0.115	0.189	0.135
MSE-t	HLN	normal	0.119	0.104	0.142	0.147	0.125	0.101	0.136	0.118
MSE-t	West	normal	0.138	0.113	0.333	0.228	0.146	0.107	0.215	0.134
MSE-t	QS	normal	0.102	0.084	0.179	0.131	0.101	0.087	0.130	0.096

Notes:
1. The data generating process is defined in equation (32), and the forecasting models are given in equations (33) and (34).
2. See the notes to Table 11.