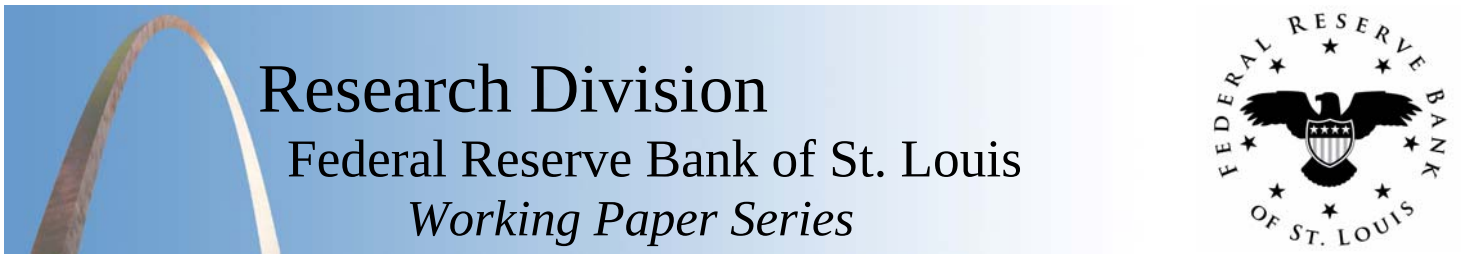

Tests of Equal Predictive Ability with Real-Time Data

Todd E. Clark
and
Michael W. McCracken

Working Paper 2008-029A
http://research.stlouisfed.org/wp/2008/2008-029.pdf

August 2008

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

Tests of Equal Predictive Ability with Real-Time Data *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Board of Governors of the Federal Reserve System

July 2008

Abstract

This paper examines the asymptotic and finite-sample properties of tests of equal forecast accuracy applied to direct, multi–step predictions from both non-nested and nested linear regression models. In contrast to earlier work in the literature, our asymptotics take account of the real-time, revised nature of the data. Monte Carlo simulations indicate that our asymptotic approximations yield reasonable size and power properties in most circumstances. The paper concludes with an examination of the real-time predictive content of various measures of economic activity for inflation.

JEL Nos.: C53, C12, C52
Keywords: Forecasting, Prediction, mean square error, causality

* *Clark (corresponding author)*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; todd.e.clark@kc.frb.org. *McCracken*: Board of Governors of the Federal Reserve System; 20th and Constitution N.W.; Mail Stop #61; Washington, D.C. 20551; michael.w.mccracken@frb.gov. We gratefully acknowledge helpful comments from Borağan Aruoba, seminar participants at Oregon, SUNY-Albany, and UBC, and participants in the following conferences: Real-Time Data Analysis and Methods at the Federal Reserve Bank of Philadelphia, Computing in Economics in Finance, International Symposium on Forecasting, NBER Summer Institute, and the ECB Workshop on Forecasting Techniques. Barbari Rossi and Juri Marcucci provided especially helpful comments in discussing the paper at, respectively, the Philadelphia Fed conference and ECB workshop. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, Board of Governors, Federal Reserve System, or any of its staff.

1 Introduction

Testing for equal out-of-sample predictive ability is a now common method for evaluating whether a new predictive model forecasts significantly better than an existing baseline model. As with in-sample comparisons (e.g. Vuong, 1989), the asymptotic distributions of the test statistics depend on whether the comparisons are between nested or non-nested models (one exception is Giacomini and White (2006)). For non-nested comparisons, Granger and Newbold (1977) and Diebold and Mariano (1995) develop asymptotically standard normal tests for predictive ability that allow comparisons between models that don't have estimated parameters. West (1996), McCracken (2000), and Corradi, Swanson and Olivetti (2001) extend these results for non-nested models to allow for estimated parameters. In these studies, the tests generally continue to be asymptotically standard normal. Chao, Corradi and Swanson (2001), Clark and McCracken (2001, 2005), Corradi and Swanson (2002), and McCracken (2007) derive asymptotics for various tests of forecasts from nested models. In much of this work, nested comparisons imply asymptotic distributions that are not asymptotically standard normal.

In this literature, one issue that is uniformly overlooked is the real-time nature of the data used in many applications. Specifically, the literature ignores the possibility that at any given forecast origin the most recent data is subject to revision. This is an issue because an out-of-sample test of predictive ability is functionally very different from an in-sample one, in a way that makes the out-of-sample test 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 $t = R, ...T,$), (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 (2006) finds), tests of predictive ability using real-time data may have a different asymptotic distribution than tests constructed using data that is never revised.

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

(2002) argue forecasts should be evaluated with real-time data because practical forecasting 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)). 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)). However, in practice, revisions are difficult to model. Koenig, Dolmas and Piger (2003) suggest instead using the various vintages of data as they would have been observed in real time to construct forecasts. More commonly, though, forecasts are generated at a moment in time using the most recent vintage of data. Accordingly, we 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.

In this paper we provide analytical, Monte Carlo and empirical evidence on pairwise tests of equal accuracy of forecasts generated and evaluated using real-time data. We consider both non-nested and nested forecast model comparisons. We restrict attention to linear direct multi-step models evaluated under quadratic loss. We also restrict attention to the case in which parameter estimates are generated on a recursive basis, with the model estimation sample growing as forecasting moves forward in time. Results for the fixed and rolling estimation schemes are qualitatively similar. As to data revisions, in some cases, we permit the revision process to consist of both "news" and "noise" as defined in Mankiw, Runkle and Shapiro (1984). In general, though, we emphasize the role of noisy revisions.

Our results indicate that data revisions can significantly affect the asymptotic behavior of tests of equal predictive ability. For example, for tests applied to forecasts from non-nested models, West (1996) shows that the effect of parameter estimation error on the test statistic can be ignored when the same loss function is used for estimation and evaluation. In the presence of data revisions, this result continues to hold only in the special case when revisions are news. When even some noise is present, parameter estimation error contributes to the asymptotic variance of the test statistic and cannot be ignored in inference.

As another example, for nested model tests of equal predictive ability, Clark and McCracken (2001, 2005) and McCracken (2007) show that standard test statistics are not asymptotically normal but instead have representations as functions of stochastic integrals. However, when the revision process contains a noise component, we show that the standard

test statistics diverge with probability one under the null hypothesis. To avoid this, we introduce a variant of the standard test statistic that is asymptotically standard normal despite being a comparison between two nested models.

The key econometric challenge to real-time analysis is that the observables are learned sequentially in time across a finite-lived revision process. For any given historical date, we therefore have multiple "observables" for a given variable. To keep our analytics as transparent as possible, while still remaining relevant for application, we assume that the revision process continues sequentially for a finite $0 \leq r << R$ periods. For a number of U.S. macroeconomic time series, such as payroll employment and the Conference Board's coincident index, this simplifying assumption is realistic. We also abstract from other forms of revisions, including benchmark revisions, leaving these complications for further research.

The paper proceeds as follows. Section 2 introduces the notation, the forecasting and testing setup, and the assumptions underlying our theoretical results. Section 3 defines the forecast tests considered, provides the null asymptotic results, and lays out how, in practice, asymptotically valid tests can be calculated. Proofs are provided in the appendix. Section 4 presents Monte Carlo results on the finite–sample performance of the asymptotics. Section 5 applies our tests to forecasts of U.S. inflation. Section 6 concludes.

2 Setup

As noted above, in our theory we allow the observables to be subject to revision over a finite number of periods, r. We have in mind the case where r is small relative to the number of observations used to estimate the model parameters at any given forecast origin. To keep track of the various vintages of a given observation we use the notation $y_s(t)$ to denote the value of the time t vintage of the observation s realization of y, where $t \geq s$. Throughout, when either there is no revision process (so that $r = 0$) or when the revision process is completed (so that $t \geq s + r$), we will drop the notation indexing the vintage and simply let $y_s(t) = y_s$. The table below illustrates our notation for a set of real-time data vintages, for a series subject to one revision, with the preliminary estimate of y in period t ($y_t(t)$) published in period t and the final estimate published in period $t + 1$ ($y_t(t + 1) = y_t$).

The sample of observations $\{\{y_s(t), x_s'(t)\}_{s=1}^{t}\}_{t=R}^{\overline{T}}$ includes a scalar random variable $y_s(t)$ to be predicted, as well as a $(k \times 1)$ vector of predictors $x_s(t)$. When the two models $i = 1, 2$ are nested we let $x_s(t) = x_{2,s}(t) = (x_{1,s}'(t), x_{22,s}'(t))'$ with $x_{i,s}(t)$ the $(k_i \times 1)$ vector

Data and vintage notation in $y_s(t)$, single revision case ($r = 1$)

data	**vintage date** (t)			
date (s)	R	$R+1$	$\cdots$	$R+P$
1	$y_1(R) = y_1$	$y_1(R+1) = y_1$	$\cdots$	$y_1(R+P) = y_1$
2	$y_2(R) = y_2$	$y_2(R+1) = y_2$	$\cdots$	$y_2(R+P) = y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$y_R(R)$	$y_R(R+1) = y_R$	$\cdots$	$y_R(R+P) = y_R$
$R+1$		$y_{R+1}(R+1)$	$\cdots$	$y_{R+1}(R+P) = y_{R+1}$
$\vdots$			$\vdots$	$\vdots$
$R+P$				$y_{R+P}(R+P)$

of predictors associated with model i. Hence the putatively nested and nesting models are linear regressions with predictors $x_{1,s}(t)$ and $x_{2,s}(t)$ respectively. When the models are non-nested we define $x_{1,s}(t)$ and $x_{2,s}(t)$ as two distinct $(k_i \times 1)$ subvectors of $x_s(t)$ (perhaps having some variables in common). Note that, throughout the paper, the definition of the sets of predictors allows for the inclusion of lagged dependent variables.

For each forecast origin t the variable to be predicted is $y_{t+\tau}(t')$, where τ denotes the forecast horizon and $t' \geq t+\tau$ denotes the vintage used to evaluate the forecasts. Throughout the evaluation period, the vintage horizon $r' = t' - t - \tau$ is fixed. At the initial forecast origin $t = R$, the vintage consists of observations (on $y_s(R)$ and $x_s(R)$) spanning $s = 1, ...R$. Letting $P-\tau+1$ denote the number of τ–step ahead predictions, the progression of forecast origins spans R through $T = R + P - \tau + 1$, each consisting of observations (on $y_s(t)$ and $x_s(t)$) spanning $s = 1, ...t$. The total number of observations in the sample corresponding to the final vintage is $\overline{T} = T + \tau + r'$. The final $\tau + r'$ vintages are used exclusively for evaluation. Under this setup, the first τ-step ahead forecast is generated using data vintage R for estimation and data vintage $R + r'$ for evaluation; the second forecast is generated with data vintage $R + 1$ and evaluated with vintage $R + r' + 1$; and so on.

Forecasts of $y_{t+\tau}(t')$, $t = R, \ldots, T$, are generated using 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), for $s = 1, ..., t-\tau$. When the models are nested, under the null hypothesis of equal forecast accuracy, model 2 nests model 1 for all t such that model 2 includes $\dim(x_{22,s}(t)) = k_{22}$ excess parameters. Then $\beta_2^* = (\beta_1^{*\prime}, 0')'$, and $y_{t+\tau}(t') - x'_{1,t}(t)\beta_1^* = u_{1,t+\tau}(t') = u_{2,t+\tau}(t') \equiv u_{t+\tau}(t')$ for all t and t'.

Both β_1^* and β_2^* are re-estimated as we progress across the vintages of data associated with each forecast origin: for $t = R, \ldots, T$, model i's $(i = 1, 2)$ prediction of $y_{t+\tau}(t')$ is

created using the parameter estimate $\hat{\beta}_{i,t}$ based on vintage t data. Models 1 and 2 yield two sequences of $P-\tau+1$ forecast errors, denoted $\hat{u}_{1,t+\tau}(t') = y_{t+\tau}(t') - x'_{1,t}(t)\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+\tau}(t') = y_{t+\tau}(t') - x'_{2,t}(t)\hat{\beta}_{2,t}$. Our assumption that the duration of the revision process is finite and small means that the parameter estimates are consistent despite the presence of "measurement error." This implies that the estimated real-time forecast errors $\hat{u}_{i,t+\tau}(t')$ are consistent for their population real-time counterparts $u_{i,t+\tau}(t')$. Moreover, all but the most recent r estimated in-sample residuals $y_{s+\tau}(t) - x'_{i,s}(t)\hat{\beta}_{i,t}$, from any given vintage, are consistent for their population counterparts $y_{s+\tau} - x'_{i,s}\beta_i^*$ based upon final revised data.

Finally, the asymptotic results below use the following additional notation. Let $h_{i,t+\tau}(t') = (y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$, $h_{i,s+\tau} = (y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$, $B_i = (Ex_{i,s}x'_{i,s})^{-1}$ and $d_{t+\tau}(t') = u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t')$. Throughout, when the models are non-nested we let $h_{t+\tau} = (h'_{1,t+\tau}, h'_{2,t+\tau})'$, $h_{t+\tau}(t') = (h'_{1,t+\tau}(t'), h'_{2,t+\tau}(t'))'$ and $U_{t+\tau} = [d_{t+\tau}(t'), h'_{t+\tau}(t') - Eh'_{t+\tau}(t'), h'_{t+\tau}, x'_t - Ex'_t]'$. When the models are nested, let $h_{t+\tau} = h_{2,t+\tau}$, $h_{t+\tau}(t') = h_{2,t+\tau}(t')$ and $U_{t+\tau} = [h'_{t+\tau}(t') - Eh'_{t+\tau}(t'), h'_{t+\tau}, x'_t - Ex'_t]'$. In either case let $H(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$. For reasons detailed below, our assumption of finite revisions allows us to define the moment matrices B_i and $H(t)$ only in terms of final data. Define the selection matrix $J' = (I_{k_1\times k_1}, 0_{k_1\times k_{22}})$ and let Ω denote the asymptotic variance of the scaled average loss differential defined more precisely in section 3.

Given the above definitions, the following assumptions (sufficient, not necessary and sufficient) are used to derive the limiting distributions in Theorems 1-3.

(A1) The parameter estimates $\hat{\beta}_{i,t}$, $i = 1, 2$, $t = R, ..., T$, are obtained by OLS for each vintage in succession and hence satisfy $\hat{\beta}_{i,t} = \arg\min_{\beta_i} \sum_{s=1}^{t-\tau}(y_{s+\tau}(t) - x'_{i,s}(t)\beta_i)^2$.

(A2) (a) $U_{t+\tau}$ is covariance stationary, (b) $EU_{t+\tau} = 0$, (c) $Ex_tx'_t < \infty$ and is positive definite, (d) For some $n > 1$ and each $j \geq 0$, $(y_t(t+j), x'_t(t+j))'$ is uniformly L^{4n} bounded, (e) $U_{t+\tau}$ is strong mixing with coefficients of size $-4n/(n-1)$, (f) Ω is positive definite.

(A3) (a) Let $K(x)$ be a kernel such that for all real scalars x, $|K(x)| \leq 1$, $K(x) = K(-x)$ and $K(0) = 1$, $K(x)$ is continuous, and $\int_{-\infty}^{\infty}|K(x)|dx < \infty$, (b) For some bandwidth M and constant $m \in (0, 0.5)$, $M = O(P^m)$.

(A4) $\lim_{R,P\to\infty} P/R = \pi \in (0, \infty)$, or (A4′) $\lim_{R,P\to\infty} P/R = 0$.

These assumptions are closely related to those in West (1996) and Clark and McCracken

(2005). Assumption 2 rules out unit roots and time trends but allows conditional heteroskedasticity and serial correlation in the levels and squares of the forecast errors. As indicated in Assumption 3, long-run variances are based on standard kernel estimators. Finally, we provide asymptotic results for situations in which the in-sample size of the initial forecast origin R and the number of predictions P are of the same order (Assumption 4) as well as when R is large relative to P (Assumption 4′).

3 Tests and Asymptotic Distributions

In this section we provide asymptotics for tests of equal forecast accuracy for non-nested and nested comparisons. For the comparison of non-nested models we allow data revisions to consist of both news and noise. In the nested case, for tractability we focus on data revisions consisting only of noise, but discuss the impact of news-only revisions. Both sets of results apply to recursive forecasts from linear models. Results for the fixed and rolling schemes differ only in the weights given to the contribution of parameter estimation error in the asymptotic variance. See West and McCracken (1998, equation 4.2) for further detail on these weights. Results for nonlinear models differ only insofar as B, $h_{t+\tau}$, and the matrix F defined below need to be redefined to account for the nonlinearity and the method used to estimate the nonlinear model. See West (1996) for further detail.

3.1 Non-nested comparisons

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal mean square error (MSE) based upon the sequence of loss differentials $\hat{d}_{t+\tau}(t') = \hat{u}^2_{1,t+\tau}(t') - \hat{u}^2_{2,t+\tau}(t')$. If we define $\text{MSE}_i = (P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{u}^2_{i,t+\tau}(t')$ $(i=1,2)$, $\bar{d} = (P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{d}_{t+\tau}(t') = \text{MSE}_1 - \text{MSE}_2$, $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t')-\bar{d})(\hat{d}_{t+\tau-j}(t'-j)-\bar{d})$, $\hat{\Gamma}_{dd}(-j) = \hat{\Gamma}_{dd}(j)$, and $\hat{S}_{dd} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dd}(j)$, the statistic takes the form

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

Under the null that the population difference in MSEs from models 1 and 2 equal zero, the authors argue that the test statistic is asymptotically standard normal.

West (1996), however, notes that this outcome depends upon whether or not the forecast errors depend upon estimated parameters. Specifically, if linear, OLS-estimated models are used for forecasting, then $P^{1/2}\bar{d} \to^d N(0,\Omega)$, where $\Omega = S_{dd} + 2(1-\pi^{-1}\ln(1+\pi))(FBS_{dh} +$

$FBS_{hh}BF'$), with $F = (-2Eu_{1,t+\tau}x'_{1,t}, 2Eu_{2,t+\tau}x'_{2,t})$, B a block diagonal matrix with block diagonal elements B_1 and B_2, S_{dd} the long-run variance of $d_{t+\tau}$, S_{hh} the long-run variance of $h_{t+\tau}$, and S_{dh} the long-run covariance of $h_{t+\tau}$ and $d_{t+\tau}$. As a result, the MSE-t test as constructed in (1) may be poorly sized because, generally speaking, the estimated variance $\hat{S}_{dd}$ is consistent for S_{dd} but not Ω.

One case in which the MSE-t test (1) will be asymptotically valid in the presence of estimated parameters is when $F = 0$. F will equal zero when the forecast error is uncorrelated with the predictors — a case that will hold when quadratic loss is used for both estimation and inference on predictive ability and the observables are covariance stationary. However, when the data is subject to revision, the population level residuals $y_{s+\tau} - x'_{i,s}\beta_i^*$, $s = 1, ..., t-\tau$, and forecast errors $y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*$, $t = R, ..., T$, need not have the same covariance structure. Consequently, $E(y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$ equaling zero need not imply anything about whether or not $E(y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$ equals zero.

Lemma 1. Define $F = 2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$ and let Assumptions 1, 2 and 4 or 4′ hold. $P^{1/2}\bar{d} = P^{-1/2}\sum_{t=R}^{T}(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t') + FBH(t)) + o_p(1)$.

The expansion in Lemma 1 is notationally identical to that in West's (1996) Lemma 4.1. This is not entirely surprising since, under our assumptions, the parameter estimates used to construct the forecasts are consistent for their population counterparts and hence the term due to parameter estimation error ($FBH(t)$) contributes in the same fashion regardless of revisions. Even so, this term is different from that in West (1996) in one very important way. The term F captures the average marginal effect of a unit change in the parameter vector $(\beta'_1, \beta'_2)'$ used to construct $\hat{d}_{t+\tau}(t')$. In the presence of revisions this is $F = 2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$. This moment need not be the same as its equivalent constructed using final revised data. Nevertheless, since the asymptotic expansion is notationally identical to West's (1996), the asymptotic distribution of the scaled average of the loss differentials remains (notationally) the same.

Theorem 1. Let Assumptions 1, 2 and 4 or 4′ hold. $P^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$ where $\Omega = S_{dd} + 2(1 - \pi^{-1}\ln(1+\pi))(FBS_{dh} + FBS_{hh}BF')$.

Since the asymptotic distribution is essentially the same as in West (1996), the special

cases in which one can ignore parameter estimation error remain essentially the same. For example, if the number of forecasts is small relative to the initial estimation sample, such that $\lim_{R,P\to\infty} P/R = \pi = 0$, then $2(1 - \pi^{-1}\ln(1+\pi)) = 0$, and hence the latter covariance terms are zero — as in West (1996).

Another special case arises when F equals zero. 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 $i = 1, 2$. Various scenarios will make $F = 0$ and permit use of conventional tests: (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; or (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.

In general, though, neither of these special cases — that $\pi = 0$ or $F = 0$ — need hold. In finite samples a small P/R doesn't guarantee that parameter estimation error is negligible, because $FBS_{dh} + FBS_{hh}BF'$ could be large. Moreover, in the presence of predictable data revisions it is typically not the case that $F = 0$. Except in the special cases noted above, generating forecasts with preliminary data — and evaluating them with either preliminary or final estimates of y — will make F non-zero, and require the standard error correction given in Theorem 1. Intuitively, what drives this result is that, in the data vintage used to estimate the models, the last small portion of observations include measurement noise (not present in most of that vintage's data sample). This noise leads to a correlation between the end-of-sample observations used to generate the forecast and the forecast error (for an error computed with preliminary or final data).

To see how real-time revisions can create a non–zero covariance between real-time forecast errors and predictors, consider a simple data-generating process (DGP) in which the final data (for t), published with a one-period delay (in $t+1$), are generated by

$$\begin{aligned} y_t &= \beta x_{1,t-1} + \beta x_{2,t-1} + e_{y,t} + v_{y,t} \\ x_{i,t} &= e_{x_i,t} + v_{x_i,t},\ i = 1, 2 \\ & \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid \text{ normal}, \end{aligned} \tag{2}$$

where e's represent innovation components that will be included in initial estimates, v's represent news components that will not, $\text{var}(e_{x_i,t}) = \sigma^2_{e,x}$, and $\text{var}(v_{x_i,t}) = \sigma^2_{v,x}$ for $i = 1, 2$

(all innovations have means of 0). Initial estimates for period t, published in t and denoted $y_t(t)$, $x_{1,t}(t)$, and $x_{2,t}(t)$, contain news and noise:

$$\begin{aligned} y_t(t) &= y_t - v_{y,t} + w_{y,t} \\ x_{i,t}(t) &= x_{i,t} - v_{x_i,t} + w_{x_i,t},\ i = 1,2 \\ & \quad w_{y,t}, w_{x_1,t}, w_{x_2,t}\ iid \text{ normal}, \end{aligned} \tag{3}$$

where v's correspond to the news component of revisions, w's denote the (mean 0) noise in the initial estimates, and $\text{var}(w_{x_i,t}) = \sigma^2_{w,x}$ $i = 1, 2$. Finally, suppose forecasts are generated from two models of the form $y_{t+1} = b_i x_{i,t} + u_{i,t+1}$, $i = 1, 2$.

The population value of the real-time forecast error for model i is

$$u_{i,t+1}(t+1) = y_{t+1}(t+1) - \beta x_{i,t}(t) = e_{y,t+1} + w_{y,t+1} + \beta x_{j,t} + \beta v_{x_i,t} - \beta w_{x_i,t}. \tag{4}$$

The noise component $w_{x_i,t}$ creates a non–zero covariance between the real-time forecast error and predictor, giving rise to a non–zero F matrix. For forecast i, this covariance is

$$\begin{aligned} E\left[u_{i,t+1}(t+1)x_{i,t}(t)\right] &= E\left[\left(e_{y,t+1} + w_{y,t+1} + \beta x_{j,t} + \beta v_{x_i,t} - \beta w_{x_i,t}\right)\left(e_{x_i,t} + w_{x_i,t}\right)\right] \\ &= -\beta\sigma^2_{w,x}. \end{aligned} \tag{5}$$

Note that, were the data generating process and forecast models to include lags of the dependent variable, noise in the dependent variable would also contribute to a non-zero covariance between the real-time forecast error and predictors. In contrast, in the complete absence of data revisions, as in West (1996), there would be no v and w terms, so the forecast error $u_{i,t+1}$ would be uncorrelated with the predictor $x_{i,t}$, and F would equal 0.

Nonetheless, in practice, even with $F \neq 0$, it is possible that a negative impact of FBS_{dh} could offset the positive impact of the variance component $FBS_{hh}BF'$. In such a setting, the correction necessitated by predictable data revisions may be small. At least in the DGPs we consider, it looks like this may often be the case. In the simple DGP above, $F = 2(\beta\sigma^2_{w,x},\ -\beta\sigma^2_{w,x})'$. If we define $\sigma^2_x = \sigma^2_{e,x} + \sigma^2_{v,x}$ and $\sigma^2_u = \sigma^2_{e,y} + \sigma^2_{v,y}$, and note that $h_{t+1} = ((\beta x_{2,t} + e_{y,t+1} + v_{y,t+1})x_{1,t}, (\beta x_{1,t} + e_{y,t+1} + v_{y,t+1})x_{2,t})'$, simple algebra yields

$$S_{hh} = \begin{pmatrix} \sigma^2_u\sigma^2_x + \beta^2\sigma^4_x & \beta^2\sigma^4_x \\ \beta^2\sigma^4_x & \sigma^2_u\sigma^2_x + \beta^2\sigma^4_x \end{pmatrix},\ S_{dh} = 2\beta\sigma^2_{e,x}\sigma^2_{e,y}[-1,1]'. \tag{6}$$

Putting together the pieces yields a population–level variance correction of

$$FBS_{dh} + FBS_{hh}BF' = \frac{8\beta^2\sigma^2_{w,x}}{\sigma^2_x}\left(\sigma^2_{w,x}\sigma^2_u - \sigma^2_{e,y}\sigma^2_{e,x}\right). \tag{7}$$

As this shows, the positive impact of noise (through $\sigma^2_{w,x}$) on the variance correction can be offset or even dominated by the negative impact associated with the information content in initial releases of y and x_1 and x_2 (through $\sigma^2_{e,y}$ and $\sigma^2_{e,x}$).

3.2 Nested comparisons

For nested models, Clark and McCracken (2005) and McCracken (2007) also propose tests for equal MSE based upon the sequence of loss differentials. Specifically, they consider the MSE-t statistic (1) and another F-type statistics, referred to as the MSE-F test.

Both tests have limiting distributions that are non–standard. Specifically, McCracken (2007) shows that, for one–step ahead forecasts from well-specified nested models, the MSE-t and MSE-F statistics converge in distribution to functions of stochastic integrals of quadratics of Brownian motion, with limiting distributions that depend on the parameter π and the number of exclusion restrictions k_{22}. For this case, simulated asymptotic critical values are provided. In Clark and McCracken (2005), the asymptotics are extended to permit direct multi-step forecasts and conditional heteroskedasticity. In this environment the limiting distributions are affected by unknown nuisance parameters. In this situation, critical values can be obtained by either Monte Carlo or bootstrap. However, all of these results are derived ignoring the potential for data revisions.

In the presence of predictable data revisions, the asymptotics for tests of predictive ability change dramatically. Again, with data revisions, the residuals $y_{s+\tau} - x'_{i,s}\beta^*_i$, $s = 1, ..., t-\tau$, and the forecast errors $y_{t+\tau}(t') - x'_{i,t}(t)\beta^*_i$, $t = R, ..., T$, need not have the same covariance structure and hence $F = 2(Eu_{2,t+\tau}(t')x'_{2,t}(t))$ need not equal zero.

Lemma 2. Define $F = 2(Eu_{2,t+\tau}(t')x'_{2,t}(t))$. Let Assumptions 1 and 2 hold and let $F(-JB_1J'+B_2) \neq 0$. (i) If Assumption 4 holds, $P^{1/2}\bar{d} = F(-JB_1J'+B_2)(P^{-1/2}\sum_{t=R}^{T} H(t))+ o_p(1)$. (ii) If Assumption 4′ holds, $R^{1/2}\bar{d} = F(-JB_1J' + B_2)(R^{1/2}H(R)) + o_p(1)$.

The expansion in Lemma 2 (i) bears some resemblance to that in Lemma 1 for non-nested models but omits the lead term $(P^{-1/2}\sum_{t=R}^{T} u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))$ because the models are nested under the null. Interestingly, neither (i) nor (ii) bears any resemblance to the corresponding expansions in Clark and McCracken (2005) and McCracken (2007) for nested models. The key reason for this difference arises from the additional assumption that $F(-JB_1J' + B_2) \neq 0$. When this condition holds, the expansion in Lemma 2 is of order

$P^{1/2}$, rather than order P as in Clark and McCracken (2005) and McCracken (2007). Not surprisingly, this implies very different asymptotic behavior of the average loss differential.

Theorem 2. Let Assumptions 1 and 2 hold and let $F(-JB_1J' + B_2) \neq 0$. (i) If Assumption 4 holds, $P^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$, where $\Omega = 2(1 - \pi^{-1}\ln(1+\pi))F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$. (ii) If Assumption 4′ holds, $R^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$, where $\Omega = F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$.

Theorem 2 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. This is in sharp contrast to the results in Clark and McCracken (2005) and McCracken (2007). This finding has a number of important implications, listed below.

1. The MSE-t test (1) diverges with probability 1 under the null hypothesis. To see this, note that by Theorem 2, the numerator of MSE-t is $O_p(1)$. Following arguments made in Clark and McCracken (2005) and McCracken (2007), the denominator of the MSE-t statistic is $O_p(P^{-1})$. Taking account of the square root in the denominator of the MSE-t test implies that the MSE-t test is $O_p(P^{1/2})$ and hence the MSE-t test has an asymptotic size of 50%. A similar argument implies the MSE-F statistic also diverges.

2. Out-of-sample inference for nested comparisons can be conducted without the strong auxiliary assumptions made in Clark and McCracken (2005) and McCracken (2007) regarding the correct specification of the models. Optimal forecasts from properly specified models will generally follow MA($\tau - 1$) processes, which we typically required in our prior work. In this paper, 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 simulation or non-standard tables. So long as an asymptotically valid estimate of Ω is available, standard normal tables can be used to conduct inference.

However, the asymptotic distribution of the MSE-t test can differ from that given in Theorem 2. The leading case occurs when the revisions consist of news rather than noise, so that $F = 0$. Another occurs when the null model is a random walk and the alternative includes variables subject to predictable revisions. Still others are listed in the discussion following Theorem 1. But even with predictable revisions that make F non-zero, Theorem 2 fails to hold when $F(-JB_1J' + B_2)$ (and hence Ω) equals zero. In both cases, the MSE-t statistic (from (1)) is bounded in probability under the null. However, in each instance the

asymptotic distributions are non-standard in much the same way as in Clark and McCracken (2005). Moreover, conducting inference using these distributions is complicated by the presence of unknown nuisance parameters. We leave a complete characterization of these distributions to future research. In the Monte Carlo section, however, we examine the ability of the distributions developed in this paper and in Clark and McCracken (2005) to approximate the more complicated, true asymptotic distributions when Ω is small.

3.3 Estimating the standard errors

To construct asymptotically valid estimates of the above standard errors, some combination of S_{dd}, S_{dh}, S_{hh}, F, B, and $2(1-\pi^{-1}\ln(1+\pi))$ needs to be estimated. Since $\hat{\pi} = P/R$ is consistent for π, estimating $\Pi \equiv 2(1-\pi^{-1}\ln(1+\pi))$ is trivial. For $\widehat{B}_i = (T^{-1}\sum_{s=1}^{T-\max(\tau,r)} x_{i,s}x_{i,s}')^{-1}$, we let $\widehat{B}$ denote the block diagonal matrix constructed using $\widehat{B}_1$ and $\widehat{B}_2$. For non-nested comparisons, we define $\widehat{F}_i = 2(-1)^i[P^{-1}\sum_{t=R}^{T}\hat{u}_{i,t+\tau}(t')x_{i,t}'(t)]$ and $\widehat{F} = (\widehat{F}_1, \widehat{F}_2)$. For nested comparisons, $\widehat{F} = 2[P^{-1}\sum_{t=R}^{T}\hat{u}_{2,t+\tau}(t')x_{2,t}'(t)]$.

For the long-run variances and covariances we consider estimates based on standard kernel-based estimators, as in West (1996), West and McCracken (1998) and McCracken (2000). To be more precise, we use kernel-weighted estimates of $\Gamma_{dd}(j) = Ed_{t+\tau}(t')d_{t+\tau-j}(t'-j)$, $\Gamma_{dh}(j) = Ed_{t+\tau}(t')h_{t+\tau-j}'$ and $\Gamma_{hh}(j) = Eh_{t+\tau}h_{t+\tau-j}'$ to estimate S_{dd}, S_{dh} and S_{hh}. To construct the relevant pieces recall that $\hat{u}_{i,t+\tau}(t') = y_{t+\tau}(t') - x_{i,t}'(t)\hat{\beta}_{i,t}$, $t = R, ..., T$. For non-nested comparisons, define $\widehat{h}_{s+\tau} = ((y_{s+\tau} - x_{1,s}'\hat{\beta}_{1,T})x_{1,s}', (y_{s+\tau} - x_{2,s}'\hat{\beta}_{2,T})x_{2,s}')'$, $s = 1, ..., T$. For nested comparisons, define $\widehat{h}_{s+\tau} = (y_{s+\tau} - x_{2,s}'\hat{\beta}_{2,T})x_{2,s}$, $s = 1, ..., T$. Let $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t')-\bar{d})(\hat{d}_{t+\tau-j}(t'-j)-\bar{d})$, $\hat{\Gamma}_{hh}(j) = T^{-1}\sum_{s=1+j}^{T}\widehat{h}_{s+\tau}\widehat{h}_{s+\tau-j}'$ and $\hat{\Gamma}_{dh}(j) = P^{-1}\sum_{t=R+j}^{T}\hat{d}_{t+\tau}(t')\widehat{h}_{t+\tau-j}'$, with $\hat{\Gamma}_{dd}(j) = \hat{\Gamma}_{dd}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}_{hh}'(-j)$ and $\hat{\Gamma}_{dh}(j) = \hat{\Gamma}_{dh}'(-j)$. Weighting the relevant leads and lags of these covariances as in Newey and West's (1987) HAC estimator, we compute the long-run variances $\hat{S}_{dd}$, $\hat{S}_{hh}$, and $\hat{S}_{dh}$. The relevant pieces are consistent for their population counterparts.

Theorem 3. Let Assumptions 1, 2 and 4 or 4′ hold. (a) $\widehat{B}_i \rightarrow^p B_i$, $\widehat{F} \rightarrow^p F$, $\hat{\Gamma}_{dd}(j) \rightarrow^p \Gamma_{dd}(j)$, $\hat{\Gamma}_{dh}(j) \rightarrow^p \Gamma_{dh}(j)$ and $\hat{\Gamma}_{hh}(j) \rightarrow^p \Gamma_{hh}(j)$. (b) If Assumption 3 holds, $\hat{S}_{dd} \rightarrow^p S_{dd}$, $\hat{S}_{dh} \rightarrow^p S_{dh}$, $\hat{S}_{hh} \rightarrow^p S_{hh}$.

Along with Theorems 1 and 2, Theorem 3 and Slutsky's Theorem imply that, given a consistent estimate of Ω, $P^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$ (or $R^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$) is asymptotically standard normal.

To estimate Ω, for non-nested comparisons one can use either $\widehat{\Omega} = \hat{S}_{dd} + 2\hat{\Pi}(\widehat{F}\widehat{B}\hat{S}_{dh} + \widehat{F}\widehat{B}\widehat{S}_{hh}\widehat{B}\widehat{F})$ or $\widehat{\Omega} = \hat{S}_{dd}$, depending on whether one expects a noise component to the data revisions. Conveniently, the former is consistent for Ω whether revisions are news or noise and hence may be a robust choice in practice. For nested comparisons, with known, noisy revisions, one can use either $\widehat{\Omega}$ =2 $\hat{\Pi}\widehat{F}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{F}'$ or $\widehat{\Omega} = \widehat{F}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{F}'$, depending upon whether or not one suspects that the $\pi > 0$ or $\pi = 0$ asymptotics are those most appropriate in a given application.

4 Monte Carlo Evidence

We proceed by first describing our Monte Carlo framework and the construction of the test statistics. We then present results on the size and power of the forecast–based tests, for forecast horizons of one and four steps. We compare results based on our proposed tests and asymptotics, which take into account the impact of data revisions, against results based on conventional tests and asymptotics, ignoring the potential impact of data revisions. The DGPs include simple ones for which we can work out analytic results and more complicated ones parameterized to roughly reflect the properties of the change in the quarterly U.S. inflation rate (y) and the output gap (x).

4.1 Monte Carlo design: non-nested case

In the non–nested forecast case, we consider three DGPs patterned broadly after those in Godfrey and Pesaran (1983). In practice, data such as GDP are subject to many revisions. In our Monte Carlo exercises, we try to simplify matters while at the same time preserving some of the essential features of actual (non-benchmark) revisions.

For DGPs 1 and 2, the final data are generated by:

$$\begin{aligned} y_t &= .4x_{1,t-1} + (.4+\beta)x_{2,t-1} + \sigma_{e,y}e_{y,t} + \sigma_{v,y}v_{y,t} \\ x_{i,t} &= \sigma_{e,x}e_{x_i,t} + \sigma_{v,x}v_{x_i,t},\ i = 1,2 \\ & \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid\ N(0,1). \end{aligned} \tag{8}$$

Across the DGP 1 and 2 experiments, the variances of the y and x variables are held fixed, but the variances of the innovation components vary, as described below. For DGP 3, the final data are generated by

$$y_t = -0.4y_{t-1} - 0.3y_{t-2} + .25x_{1,t-1} + (.25+\beta)x_{2,t-1} + \sigma_{e,y}e_{y,t} + \sigma_{v,y}v_{y,t} \tag{9}$$

$$
\begin{aligned}
x_{i,t} &= 1.1x_{i,t-1} - 0.3x_{i,t-2} + \sigma_{e,x}e_{x_i,t} + \sigma_{v,x}v_{x_i,t},\ i = 1,2 \\
& \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid\ N(0,1).
\end{aligned}
$$

For all DGPs, the coefficient β is set to zero in size experiments. In power experiments, β is set to 0.6 in DGPs 1 and 2 and 0.75 in DGP 3.

For the revision process, we assume a single revision of an initially published estimate. For analytical tractability, in DGPs 1 and 2 the final values are published with just a one-period delay. In DGP 3, the final values are published with a four-period delay. Specifically, a first estimate of each variable's value in period t is published in period t (denoted $y_t(t)$, $x_{1,t}(t)$, and $x_{2,t}(t)$). The final estimates (y_t, $x_{1,t}$, and $x_{2,t}$) are treated as being published in period $t+1$ in DGPs 1 and 2 and period $t+4$ in DGP 3.

In light of evidence of predictability in data revisions (e.g., Croushore and Stark (2003), Faust and Wright (2005), and Arouba (2006)), the revision processes have a common general structure that incorporates unpredictable (news) and unpredictable (noise) components:

$$
\begin{aligned}
y_t(t) &= y_t - \sigma_{v,y}v_{y,t} + \sigma_{w,y}w_{y,t} \qquad (10) \\
x_{i,t}(t) &= x_{i,t} - \sigma_{v,x}v_{x_i,t} + \sigma_{w,x}w_{x_i,t},\ i = 1,2 \\
& \quad w_{y,t}, w_{x_1,t}, w_{x_2,t}\ iid\ N(0,1)
\end{aligned}
$$

With this structure, the initial estimates include all of the information in the final value except the innovation (news) components denoted by v's (incorporated in, e.g., $y_t - \sigma_{v,y}v_{y,t}$), but add in measurement error (noise) components denoted by w's.

For DGPs 2 and 3, our parameterizations of the revision processes are roughly drawn from evidence in Aruoba (2006) and empirical estimates for real-time U.S. data on the change in GDP inflation and the HP output gap from 1965 through 2003. For DGP 1, however, we use a parameterization designed to yield a more sizable impact of data revisions on real-time forecast inference: $\sigma^2_{e,y} = 0.1$, $\sigma^2_{v,y} = 0.9$, $\sigma^2_{w,y} = 0.2$, $\sigma^2_{e,x} = 1.7$, $\sigma^2_{v,x} = .3$, and $\sigma^2_{w,x} = 2$. Under this parameterization, the correlation of the revision in y with the initial estimate is about -0.2, in line with our data. However, the revision variance is nearly 70 percent of the variance of y, well above the 30 percent average reported by Aruoba (2006). The correlation of the revision in each x variable with the initial estimate is nearly -0.7, a bit higher than in actual data for the output gap. But the revision variance of each x variable is 15 percent larger than the variance of the corresponding final series.

In DGP 2 experiments, we use $\sigma^2_{e,y} = 0.8$, $\sigma^2_{v,y} = .2$, $\sigma^2_{w,y} = 0.2$, $\sigma^2_{e,x} = 1.7$, $\sigma^2_{v,x} = 0.3$,

and $\sigma^2_{w,x} = 0.5$. In DGP 2, the correlation of the revision in y with the initial estimate remains around -0.2, roughly in line with our data. The correlation of the revision in the x variables with the initial estimate is nearly -0.4, less than in our actual data on the output gap, but not out of line with evidence for other variables. As a share of the variance of the final data, the variance of revisions is about 20 percent for y and 40 percent for the x variables. These settings balance evidence from our data with the broader results in Aruoba (2006). Finally, we parameterize DGP 3 to obtain magnitudes of revisions and predictability in line with DGP 2: $\sigma^2_{e,y} = 0.8$, $\sigma^2_{v,y} = 0.2$, $\sigma^2_{w,y} = 0.2$, $\sigma^2_{e,x} = 0.2$, $\sigma^2_{v,x} = 0.3$, and $\sigma^2_{w,x} = 0.5$.

With DGPs 1 and 2, we test for equal accuracy of τ-horizon forecasts from models

$$y^{(\tau)}_{t+\tau} = a_1 x_{1,t} + u_{1,t+\tau} \tag{11}$$

$$y^{(\tau)}_{t+\tau} = b_1 x_{2,t} + u_{2,t+\tau}, \tag{12}$$

where $y^{(\tau)}_{t+\tau} \equiv \tau^{-1} \sum_{s=1}^{\tau} y_{t+s}$. The forecasting models for DGP 3 experiments take the form:

$$y^{(\tau)}_{t+\tau} = a_0 + a_1 y_t + a_2 y_{t-1} + a_3 x_{1,t} + u_{1,t+\tau} \tag{13}$$

$$y^{(\tau)}_{t+\tau} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 x_{2,t} + u_{2,t+\tau}. \tag{14}$$

At each forecast origin t, the observable time series for each variable consists of initial or first vintage estimates for period $t - r + 1$ through t and final values for periods $t - r$ and earlier, where $r = 1$ in DGPs 1 and 2 and $r = 4$ in DGP 3. As forecasting moves forward, the models are recursively re-estimated with an expanding sample of data, by OLS.

In evaluating forecasts, we compute forecast errors using actual values of $y_{t+\tau}$ taken to be the initial estimate published in period $t + \tau$, $y_{t+\tau}(t + \tau)$. We form two versions of the MSE-t test, one with a standard error of just $\widehat{S}_{dd}$ and the other with $\widehat{\Omega} = \widehat{S}_{dd} + 2\widehat{\Pi}(\widehat{F}\widehat{B}\widehat{S}_{dh} + \widehat{F}\widehat{B}\widehat{S}_{hh}\widehat{B}\widehat{F}')$. For experiments in which data revisions are purely news, the variance correction incorporated in $\widehat{\Omega}$ is not necessary, but asymptotically irrelevant — so either test is valid. For experiments in which data revisions include noise, the variance correction incorporated in $\widehat{\Omega}$ is necessary. With noisy revisions, the test based on $\widehat{\Omega}$ is valid, while the test based on just $\widehat{S}_{dd}$ will (asymptotically) be inaccurate.

We estimate the long-run variances $\widehat{S}_{dd}$, $\widehat{S}_{dh}$, and $\widehat{S}_{hh}$ as in Newey and West (1987), using a bandwidth of 2τ. This bandwidth setting allows for noise in data revisions to create some serial correlation in even one–step ahead forecast errors. For example, if the true model is an AR(1), with revisions as in (10), one–step ahead real-time forecast errors

will contain an MA(1) component, due to noise. Our use of a constant bandwidth follows common practice (e.g., Orphanides and van Norden (2005)). However, like Newey and West (1987), our asymptotics require that the bandwidth increase with sample size. Were we to widely vary the sample size, we would want to depart from the fixed-bandwidth approach.

All test statistics are compared against critical values from the standard normal distribution. We report the percentage of 10,000 simulations in which the null of equal accuracy is rejected at the 5% significance level. Finally, with quarterly data in mind, we consider a range of sample sizes: $P = 20$, 40, 80, and 160. For simplicity, we report results for a single R setting, of 80; results with $R = 40$ are very similar.

4.2 Monte Carlo design: nested case

In the nested forecast case, we also consider three DGPs. The DGPs include both news and noise components of revisions, and we consider alternative parameterizations that make our proposed correction important or not.

For DGPs 1 and 2, the final data are generated by

$$\begin{aligned} y_t &= .7y_{t-1} + \beta_{22}x_{t-1} + e_{y,t} + v_{y,t} \qquad (15) \\ x_t &= .7x_{t-1} + e_{x,t} + v_{x,t} \\ Var\begin{pmatrix} e_{y,t} \\ e_{x,t} \\ v_{y,t} \\ v_{x,t} \end{pmatrix} &= \begin{pmatrix} .8 & & & \\ cov(e_y, e_x) & 0.2 & & \\ 0 & 0 & 0.2 & \\ 0 & 0 & 0 & 0.3 \end{pmatrix} \end{aligned}$$

In DGP 1 experiments, $cov(e_y, e_x) = 0.35$; in DGP 2, $cov(e_y, e_x) = 0.25$. The DGPs also differ in their parameterizations of the noise process, as described below. In size experiments, $\beta_{22} = 0$; in power experiments, $\beta_{22} = 0.3$.

For DGP 3, the final data are generated by

$$\begin{aligned} y_t &= -0.4y_{t-1} - 0.3y_{t-2} - 0.2y_{t-3} + .1y_{t-4} + \beta_{22}x_{t-1} + e_{y,t} + v_{y,t} \qquad (16) \\ x_t &= 1.1x_{t-1} - 0.3x_{t-2} + e_{x,t} + v_{x,t} \\ Var\begin{pmatrix} e_{y,t} \\ e_{x,t} \\ v_{y,t} \\ v_{x,t} \end{pmatrix} &= \begin{pmatrix} 0.8 & & & \\ 0.25 & 0.2 & & \\ 0 & 0 & 0.2 & \\ 0 & 0 & 0 & 0.3 \end{pmatrix} \end{aligned}$$

In size experiments, $\beta_{22} = 0$; in power experiments, $\beta_{22} = 0.3$.

For all DGPs, the revision processes take the form:

$$y_t(t) = y_t - v_{y,t} + w_{y,t} \qquad (17)$$

$$
\begin{aligned}
x_t(t) &= x_t - v_{x,t} + w_{x,t} \\
& \\
& w_{y,t}, w_{x,t} \; iid \text{ normal},
\end{aligned}
$$

Final data are released with delays of one (DGPs 1 and 2) or four (DGP 3) periods.

In DGP 1, the noise innovation variances are set to $\sigma^2_{w,y} = 1.8$ and $\sigma^2_{w,x} = 0.5$. Under this parameterization, the correlation of the revision in y with the initial estimate is about -0.7, and the variance of revisions is about the same as the variance of the final data on y — well above what is observed in data on inflation. With $cov(e_y, e_x) = 0.35$, this parameterization yields a relatively large Ω, taken as the baseline case. DGPs 2 and 3 use $\sigma^2_{w,y} = .2$ and $\sigma^2_{w,x} = 0.5$, which makes the correlation of the revision in y with the initial estimate about -0.25, and the variance of revisions in y about 20 to 30 percent of the variance of the final data on y — roughly in line with actual data. Analytically, we have verified that the DGP 2 parameterization using $cov(e_y, e_x) = 0.25$ yields a relatively small population Ω.

In DGP 1 and 2 experiments, we test for equal accuracy of τ-horizon forecasts from

$$
\begin{aligned}
y^{(\tau)}_{t+\tau} &= a_1 y_t + u_{1,t+\tau} && (18) \\
y^{(\tau)}_{t+\tau} &= a_2 y_t + b_2 x_t + u_{2,t+\tau}, && (19)
\end{aligned}
$$

where $y^{(\tau)}_{t+\tau} \equiv \tau^{-1} \sum_{s=1}^{\tau} y_{t+s}$. In DGP 3 experiments, the forecasting models are

$$
\begin{aligned}
y^{(\tau)}_{t+\tau} &= a_0 + a_1 y_t + a_2 y_{t-1} + a_3 y_{t-2} + a_4 y_{t-3} + u_{1,t+\tau} && (20) \\
y^{(\tau)}_{t+\tau} &= b_0 + b_1 y_t + b_2 y_{t-1} + b_3 y_{t-2} + b_4 y_{t-3} + b_5 x_t + u_{2,t+\tau}. && (21)
\end{aligned}
$$

At each forecast origin t, the observable time series for each variable consists of an initial or first vintage estimates for period $t - r + 1$ through t and final values for periods $t - r$ and earlier, where $r = 1$ in DGPs 1 and 2 and $r = 4$ in DGP 3. The parameters of the forecasting models are estimated recursively by OLS.

In evaluating forecasts, we compute forecast errors using actual values of $y_{t+\tau}$ taken to be the initial estimate published in period t, $y_{t+\tau}(t + \tau)$. The null hypothesis is that the variables included in the larger model and not the smaller have no predictive content. We consider various tests of equal MSE and the Chao, Corradi, and Swanson (2001) test of out-of-sample Granger causality (henceforth, the CCS test). Two of the tests take into account the impact of noisy data revisions. The first is our MSE-$t(\Omega)$ test, using the square root of $\widehat{\Omega} = 2\widehat{\Pi}\widehat{F}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{F}'$ as the standard error, which we compare against standard normal critical values. The second is the CCS test, compared against χ^2

critical values. We construct the CCS test to account for data revisions simply by using the real-time forecast errors and predictors in computing the moments entering the variance given in Chao, Corradi, and Swanson (2001).

We consider several other tests that ignore the potential impact of data revisions, based primarily on the asymptotics of Clark and McCracken (2005). Specifically, we construct the MSE-F test and compare it against asymptotic critical values simulated as in Clark and McCracken (2005). We also construct the conventional version of the MSE-t test, defined as MSE-$t(S_{dd}) = P^{1/2}(MSE_1 - MSE_2)/\widehat{S}_{dd}^{1/2}$, and compare it against both Clark and McCracken (2005) and standard normal critical values. For each test, we reject the null if the test statistic exceeds the relevant 5% critical value (taken from the right tail, in the case of the MSE tests). We compute $\widehat{S}_{dd}$, $\widehat{S}_{hh}$, and the long-run variances entering the CCS test with Newey and West's (1987) HAC estimator, using 2τ lags.

Our various DGP parameterizations will likely affect the performances of the tests that take revisions into account versus those that do not. With DGP 1, parameterized to make Ω large, Theorem 2 implies our proposed test to be the correct one, while the MSE-F and MSE-$t(S_{dd})$ statistics compared against Clark and McCracken (2005) critical values should reject far too often. For DGPs 2 and 3, parameterized to make Ω small, our proposed test remains asymptotically valid, while the Clark-McCracken tests that ignore data revisions are not. However, with Ω small, it is possible that, practically speaking, the Clark and McCracken (2005) asymptotics will be as good as this paper's asymptotics that account for revisions. In unreported results, we have also considered versions of DGPs 1 and 2 with noisy revisions but parameterized to make $\Omega = 0$. In those cases, neither this paper's asymptotic approximation (Theorem 2) nor the asymptotic results of our prior work formally apply. These results are very similar to those for the small Ω case.

4.3 Monte Carlo results: non-nested case

Table 1 reports size and power results from non–nested forecast simulations. We first consider the properties of forecast tests in which revisions contain only news, in which case $F = 0$, and no standard error correction is needed. We then consider results for noisy revisions, in which case $F \neq 0$ and, in principle, a standard error correction is necessary.

With revisions that contain only news, the size of the unadjusted t-test for equal MSE (MSE-$t(S_{dd})$) ranges from 6 to 28 percent — such that the test ranges from slightly to significantly oversized. With large forecast samples, the test tends to be close to correctly

sized. But the size of the test rises as the sample shrinks and as the horizon increases.

Even though no standard error correction is asymptotically necessary, the adjusted t-test (MSE-$t(\Omega)$) seems to have better small-sample size properties, at least in smaller forecast samples. Across results for DGPs 1-3 with just news, the size of the adjusted test ranges from 4 to 14 percent — from slightly undersized to modestly oversized. These results indicate that, in a context in which a practitioner can't be sure that the revisions in his/her data set contain only news and not noise, applying our correction won't harm inference if the revisions contain only news, and may improve it.

Consider now the size of the tests in the case of predictable revisions (noise). In this case, the unadjusted MSE-t test might be expected to be oversized, more so for larger P/R than smaller P/R, because the variance in the test fails to account for the variance impact of the predictable revisions. In the case of one–step ahead forecasts from DGPs 1 and 2, we can analytically compute the population value of the correction $FBS_{dh} + FBS_{hh}BF'$ from (7), to be 2.34 for DGP 1 and -0.28 for DGP 2, compared to population S_{dd} values of roughly 3.5. Accordingly, based on the one–step ahead asymptotics, we might expect the unadjusted and adjusted tests to both be about correctly sized in results for DGP 2, but the unadjusted to be oversized in results for DGP 1.

The noisy revisions results in Table 1 are consistent with these asymptotics. In DGP 1 results, the unadjusted test is modestly to significantly oversized, yielding a rejection rate between 11 and 12 percent at the one-step horizon and between 8 and 23 percent at the four-step horizon. The adjusted test has much better properties, ranging from slightly undersized to modestly oversized: the rejection rates range from 4 to 6 percent for one-step forecasts and 3 to 9 percent for four-step forecasts. In DGP 2 results, the size of the unadjusted test is consistently lower, sometimes significantly, while the size of the adjusted test is typically a bit higher than in the DGP 1 results. With DGP 2, the size of the unadjusted test ranges from 5 to 9 percent at the one-step horizon and from 5 to 21 percent at the four-step horizon. The size of the adjusted test falls between 5 and 8 percent for one-step forecasts and 3 and 11 percent for four-step forecasts. Results for DGP 3, for which the standard error correction is also quite small, are qualitatively similar to those for DGP 2. Overall, the results for the noise revisions simulations are similar to those for the news case in that, in small samples, the adjusted t-test generally has better small sample properties than the unadjusted test, even if, in population, the necessary correction is small.

The lower half of Table 2 shows that, with revisions that contain only news, the powers of the adjusted and unadjusted tests for one-step ahead forecasts are virtually identical. For example, with DGP 3, $P = 80$, and $\tau = 1$, both tests have power of 80 percent. At the four-step horizon, power is generally lower, and, for smaller samples, the power of the unadjusted test is often modestly greater than that of the adjusted test. For instance, with DGP 1, $P = 40$, and $\tau = 4$, the powers of the unadjusted and adjusted tests are 24 and 19 percent, respectively. With predictable revisions (noise) in all variables, differences in power remain small or modest for small P, but overall power can be trivial. In DGPs 1 and 3, the power of the unadjusted test ranges from 8 to 24 percent; the power of the adjusted test ranges from 4 to 17 percent. Power is much better for DGP 2, with trivial differences across the unadjusted and adjusted tests, except with small P and the four-step horizon. With DGP 2, the powers of one–step ahead tests fall between 22 and 84 percent; the powers of the four-step tests range from 12 to 23 percent.

4.4 Monte Carlo results: nested case

Table 2 reports size and power results from nested model forecast simulations. Consistent with our theoretical results, with DGP 1 (for which we can analytically determine that Ω is large), the standard MSE-F and MSE-$t(S_{dd})$ statistics compared against Clark and McCracken (2005) critical values suffer large size distortions. The size of the MSE-F test ranges from 5 to 26 percent; the size of the MSE-$t(S_{dd})$ test ranges from 21 to 32 percent. Comparing MSE-$t(S_{dd})$ against standard normal critical values also yields significant oversizing, with size ranging from 10 to 24 percent. Comparing our proposed statistic MSE-$t(\Omega)$ against standard normal critical values yields much more accurate inference, with size between 8 and 10 percent at the one-step horizon and between 13 and 15 percent at the four-step horizon. Admittedly, at the four-step horizon, all of the MSE tests are oversized; however, our proposed test fares at least slightly better than the others. The CCS test, which also accounts for the distributional impact of data revisions, is reasonably accurate for one-step ahead forecasts and long samples of four-step ahead forecasts (with size of between 5 and 7 percent), but oversized for smaller samples of four-step ahead forecasts (with size between 9 percent for $P = 80$ and 21 percent for $P = 20$).

With DGPs 2 and 3, for which Ω is small, it is less clear that one test is more reliable than the others. The MSE-F test seems most reliable, with size ranging from 4 to 10 percent. The MSE-$t(S_{dd})$ test compared against Clark and McCracken (2005) critical values

is less reliable (mostly so for small P or four-step forecasts), with size between 5 and 21 percent. Comparing the same test against standard normal critical values tends to yield an undersized test (except for small P and longer forecast horizons). Our proposed MSE-$t(\Omega)$ test is consistently oversized, with size between 6 and 14 percent. Finally, the CCS test is correctly sized or modestly oversized for one-step ahead forecasts and long samples of four-step ahead forecasts, but oversized for smaller samples of four-step ahead forecasts. Experiments with news-only revisions or noise revisions but $\Omega = 0$ yield similar results.

The lower half of Table 2 provides power results from nested model simulations. With DGP 1, the MSE-F and MSE-$t(\Omega)$ tests have comparable power: e.g., at the one-step horizon, the power of the former ranges from 39 to 98 percent; the power of the latter ranges from 63 to 93 percent. Comparing the conventional MSE-$t(S_{dd})$ statistic against Clark and McCracken (2005) or standard normal critical values often yields lower power, more so with normal critical values. For instance, at the one-step horizon, comparing MSE-$t(S_{dd})$ against standard normal critical values yields a rejection rate between 28 and 88 percent. In experiments for DGPs 2 and 3, power is generally much lower. At the one-step horizon, the power of the MSE-F test ranges from 23 to 48 percent; the power of the MSE-$t(\Omega)$ test varies from 23 to 40 percent. In the same experiments, the power of MSE-$t(S_{dd})$ compared against standard normal critical values falls between 4 and 9 percent. The CCS test shows a reversed pattern: power is very low for DGP 1, and somewhat better for the other DGPs. Overall, the CCS test generally (although not universally, especially for large P) has lower power than the MSE-F and MSE-$t(\Omega)$ tests, at least partly because it is two-sided instead of one-sided.

On balance, in the face of potentially predictable data revisions in nested model forecast comparisons, it would seem useful to consider results from multiple tests, preferably MSE-F and MSE-$t(\Omega)$. In cases in which Ω is large, our proposed test MSE-$t(\Omega)$ should be preferred, but in many practical settings, Ω seems likely to be small. In such settings, the MSE-F test compared against Clark and McCracken (2005) critical values — which is technically valid only in the absence of revisions — still seems to work reasonably despite the potential impact of revisions on the asymptotic distribution.

5 Application to Inflation Forecasting

We apply the tests and inference approaches described above to determine whether, in real-time data (as in, e.g., Giacomini and Rossi (2005) and Orphanides and van Norden (2005)), various measures of economic activity have predictive content for inflation.

5.1 Data

Data on GDP and the GDP price index are taken from the Federal Reserve Bank of Philadelphia's Real–Time Data Set for Macroeconomists (RTDSM). The full forecast evaluation period runs from 1970:Q1 through 2003:Q4. For each forecast origin t in 1970:Q1 through 2003:Q4, we use data vintage t to estimate output gaps, (recursively) estimate the forecast models, and then construct forecasts for periods t and beyond. The starting point of the model estimation sample is always 1961:4. 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 periods.

5.2 Models

Following Stock and Watson (2003), among others, we obtain forecasts of the change in inflation at the one-year-ahead horizon from reduced–form Phillips curves:

$$\pi_{t+4}^{(4)} - \pi_t = \alpha_0 + \sum_{l=0}^{3} \alpha_l \Delta\pi_{t-l} + \beta x_t + u_{PC,t+4}, \tag{22}$$

where inflation is $\pi_t^{(4)} \equiv 100 \ln(p_t/p_{t-4})$, $\pi_t^{(1)} \equiv \pi_t$, and x_t is a measure of economic activity. In one model, x_t is defined as the four–quarter GDP growth rate, $\ln(\text{GDP}_t/\text{GDP}_{t-4})$. In another model, x_t is defined as HP-detrended log GDP.

In addition to comparing forecasts from one version of (22) with GDP growth to another with the output gap, we compare forecasts from the model with GDP growth to forecasts from the following AR model for the change in inflation:

$$\pi_{t+4}^{(4)} - \pi_t = \alpha_0 + \sum_{l=0}^{1} \alpha_l \Delta\pi_{t-l} + u_{AR,t+4}. \tag{23}$$

In computing the MSE-t and CCS tests, we use the Newey and West (1987) estimator of the necessary long-run variances, with a bandwidth of 8.

5.3 Results

The top panel of Table 3 presents results for the (non–nested) comparison of forecasts from the models with the output gap (model 1) and GDP growth (model 2). For all samples and definitions of actuals, the model with GDP growth yields more accurate forecasts. However, there is little evidence of statistical significance in the forecast accuracy differences. If the conventional variance $\widehat{S}_{dd}$ is used in forming the t-test, the null of equal accuracy is rejected for all 1985-2003 samples. However, consistent with our Monte Carlo evidence, taking account of the potential for predictability in the data revisions raises the estimated standard error. All rejections of equal accuracy based on the t-test using $\widehat{S}_{dd}$ go away when the test uses the adjusted variance $\widehat{\Omega}$.

The bottom panel provides results for the (nested) comparison of forecasts from the AR(2) model (model 1) and the model with four lags of inflation and GDP growth (model 2). For nearly all samples and definitions of actuals, the forecasts from the model with GDP growth are more accurate than the AR(2) forecasts. 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 always reject the null AR model in the 1970-2003 and 1970-1984 samples. Taking account of data revisions by using the variance $\widehat{\Omega}$ in the MSE-t test always increases the (absolute) value of the t–statistic — but in only one case is the adjusted t–statistic significant when the unadjusted t–statistic (compared against standard normal critical values) is not. Overall, the MSE-F and MSE-$t(\Omega)$ tests, most reliable in the Monte Carlo evidence, always agree.

6 Conclusion

This paper derives the limiting distributions for tests of equal predictive ability based on real-time data. When revised data are used to construct and evaluate forecasts, tests of equal MSE typically do not have the same asymptotic distributions as when the data is never revised. However, suitably modified tests are asymptotically standard normal and hence inference can be conducted using the relevant tables. Monte Carlo simulations broadly confirm our asymptotic approximations. Taking revisions into account by using our proposed tests rather than the standard forms of the tests can yield more reliable inferences, although in practice, there will be situations in which our proposed corrections are not very important. We conclude by applying our tests to competing forecasts of U.S. inflation.

References

Aruoba, S.B. (2006), "Data Revisions Are Not Well-Behaved," *Journal of Money, Credit, and Banking*, forthcoming.

Chao, J., Corradi, V., Swanson, N.R. (2001), "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598-620.

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

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

Clark, T.E., and McCracken, M.W. (2007), "Tests of Equal Predictive Ability with Real-Time Data," Research Working Paper 07-06, Federal Reserve Bank of Kansas City.

Corradi, V., and Swanson, N.R. (2002), "A Consistent Test for Nonlinear Out–of–Sample Predictive Accuracy," *Journal of Econometrics*, 110, 353-381.

Corradi, V., Swanson, N.R., and Olivetti, C. (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 105, 315-358.

Croushore, D. (2006), "Forecasting with Real–Time Macroeconomic Data," in *Handbook of Economic Forecasting* , eds. G. Elliott, C. Granger, and A. Timmermann, Amsterdam: North-Holland, pp. 961-982.

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

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

Faust, J., and Wright, J.H. (2005), "News and Noise in G-7 GDP Announcements," *Journal of Money, Credit, and Banking*, 37, 403-420.

Giacomini, R., and Rossi, B. (2005), "Detecting and Predicting Forecast Breakdowns," manuscript, Duke University.

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

Godfrey, L.G., and Pesaran, M.H. (1983), "Tests of Non–Nested Regression Models: Small Sample Adjustments and Monte Carlo Evidence," *Journal of Econometrics*, 21, 133-154.

Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, New York: Academic Press.

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

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

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

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

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

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

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

Stark, T., and Croushore, D. (2002), "Forecasting with a Real-Time Data Set for Macroeconomists," *Journal of Macroeconomics*, 24, 507-531.

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

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

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

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

7 Appendix 1: Theory Details

Most results follow from very similar arguments to those in West (1996), McCracken (2000), and Clark and McCracken (2005) but keeping track of the fact that while $(x'_{i,t}, y_{t+\tau})'$ is covariance stationary, it need not have the same first and second moments as $(x'_{i,t}(t), y_{t+\tau}(t'))'$ due to the revision process.

In order to keep track of this distinction, some notation is useful: $B_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}x'_{i,s})^{-1}$, $\hat{B}_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}(t)x'_{i,s}(t))^{-1}$, $G_i(t) = t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}y_{s+\tau}$, $\hat{G}_i(t) = t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}(t)y_{s+\tau}(t)$, $H_i(t) = t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$, and $\hat{H}_i(t) = t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau}(t) - x'_{i,s}(t)\beta_i^*)x_{i,s}(t)$. If we let $\hat{H}_i(t) - H_i(t) = t^{-1}v_{i,t}$ we obtain the identity $\hat{\beta}_{i,t} \equiv \hat{B}_i(t)\hat{G}_i(t) = \beta_i^* + B_i(t)H_i(t) + B_i(t)(t^{-1}v_{i,t}) + (\hat{B}_i(t) - B_i(t))G_i(t) + (\hat{B}_i(t) - B_i(t))(t^{-1}v_{i,t})$. In addition, we let $\sup_t$ denote $\sup_{R\le t\le T}$, and for any matrix A with elements $a_{i,j}$, $|A|$ denotes $\max_{i,j}|a_{i,j}|$. Finally, ignoring the finite sample distinction between summing over P and $P-\tau+1$ elements, each set of results are based upon the decomposition of $\bar{d}$ into four bracketed $\{.\}$ terms.

$$
\begin{aligned}
\bar{d} &= \{P^{-1}\textstyle\sum_{t=R}^{T}(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(2\sum_{i=1}^{2}(-1)^i h_{i,t+\tau}(t')B_i(t)H_i(t))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(\sum_{i=1}^{2}(-1)^{i+1}H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(\sum_{i=1}^{2}(-1)^i(2h_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t}) + 2h_{i,t+\tau}(t')(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)(t^{-1}v_{i,t}) - 2H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- (t^{-1}v'_{i,t})B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)(t^{-1}v_{i,t}) - 2(t^{-1}v'_{i,t})B_i(t)x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- \hat{H}'_i(t)(\hat{B}_i(t) - B_i(t))x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t)))\}
\end{aligned} \tag{24}
$$

Proof of Lemma 1: Given the decomposition in (24), it suffices to show that the $P^{1/2}$-scaled second bracketed term equals $FB(P^{-1/2}\sum_{t=R}^{T}H(t)) + o_p(1)$ and the $P^{1/2}$-scaled third and fourth bracketed terms are $o_p(1)$. That the first term equals $FB(P^{-1/2}\sum_{t=R}^{T}H(t)) + o_p(1)$ follows from algebra nearly identical to that in West (1996, Lemma 4.1; see apx. Lemma A4). Proofs that the third term, and each component of the fourth term are $o_p(1)$ follow similar logic. For example,

$$
\begin{aligned}
&|P^{-1/2}\textstyle\sum_{t=R}^{T}H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t)| \le \\
&k^8(P^{-1/2})(P^{-1}\textstyle\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)|)(\sup_t|B_i(t)|)^2(\sup_t|P^{1/2}H_i(t)|)^2
\end{aligned}
$$

and

$$
\begin{aligned}
&|P^{-1/2}\textstyle\sum_{t=R}^{T}h'_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t})| \le \\
&k^2(P^{1/2}/R)(\sup_t|B_i(t)|)(P^{-1}\textstyle\sum_{t=R}^{T}|h_{i,t+\tau}(t')||v_{i,t}|)
\end{aligned}.
$$

Since Assumption 2 suffices for $P^{-1}\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)|$, $\sup_t|B_i(t)|$, $\sup_t|P^{1/2}H_i(t)|$, and $P^{-1}\sum_{t=R}^{T}|h_{i,t+\tau}(t')||v_{i,t}|$ to each be $O_p(1)$, Assumption 4 or 4′ imply each term is $o_p(1)$.

Proof of Theorem 1: Given Lemma 1 the result follows nearly identical logic to that in West (1996) Theorem 4.1.

Proof of Lemma 2: First note that, under the null, the initial bracketed term in (24) is zero since $u_{1,t+\tau}(t') = u_{2,t+\tau}(t') = u_{t+\tau}(t')$. (i) If we also note that $J'x_{2,t}(t) = x_{1,t}(t)$ so that $J'H_2(t) = H_1(t)$, and take account of the (nested model) definition of F, the proof is identical to that in Lemma 1.

(ii) The key differences in the proof are (a) the scaling by $R^{1/2}$ rather than $P^{1/2}$, and (b) given our assumptions, $\sup_t R^{1/2}|B_i(t)H_i(t) - B_i(R)H_i(R)| = o_p(1)$; a result that follows from Lemma 8 of Clark and McCracken (2001). With this tool in hand we first show that the second term in (26) equals $F(-JB_1J' + B_2)(R^{1/2}H(R)) + o_p(1)$. To do so note that

$$\begin{array}{l} R^{1/2}P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')B_i(t)H_i(t) = R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t'))B_i(R)H_i(R) \\ +R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')(B_i(t)H_i(t) - B_i(R)H_i(R))) \end{array} .$$

Since $B_i(R) \to_p B_i$, $P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t') \to_p Eh'_{i,t+\tau}(t')$, $R^{1/2}H_i(R) = O_p(1)$, and

$$\begin{array}{l} |R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')(B_i(t)H_i(t) - B_i(R)H_i(R)))| \leq \\ k(P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')|)(\sup_t R^{1/2}|B_i(t)H_i(t) - B_i(R)H_i(R)|) \end{array}$$

we obtain the desired result.

Proofs that the third term, and each component of the fourth term are $o_p(1)$ follow logic comparable to that in Lemma 1 but adjusting for the rescaling. Using the same examples from Lemma 1,

$$\begin{array}{l} |R^{1/2}P^{-1}\sum_{t=R}^{T} H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t)| \leq \\ k^4(R^{-1/2})(P^{-1}\sum_{t=R}^{T} |x_{i,t}(t)x'_{i,t}(t)|)(\sup_t |B_i(t)|)^2(\sup_t |R^{1/2}H_i(t)|)^2 \end{array}$$

and

$$\begin{array}{l} |R^{1/2}P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t})| \leq \\ k(R^{-1/2})(\sup_t |B_i(t)|)(P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')||v_{i,t}|) \end{array} .$$

Since Assumption 2 suffices for $P^{-1}\sum_{t=R}^{T} |x_{i,t}(t)x'_{i,t}(t)|$, $\sup_t |B_i(t)|$, $\sup_t |R^{1/2}H_i(t)|$, and $P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')||v_{i,t}|$ to each be $O_p(1)$, Assumption 4′ implies each term is $o_p(1)$.

Proof of Theorem 2: (i) Given Lemma 2 (i) the result follows nearly identical logic to that in West (1996) Theorem 4.1. (ii) Given Lemma 2 (ii), and the fact that $R^{1/2}H(R) \to_d N(0, S_{hh})$, the result is immediate.

Proof of Theorem 3: (i) Assumption 2 suffices for $\hat{B}_i \to_p B_i$. That $\hat{F} \to_p F$ follows nearly identical arguments to that in Lemma 5.1 of West (1996). That $\hat{\Gamma}_{hh} \to_p \Gamma_{hh}$ is immediate from Theorem 5.1 of West (1996). As detailed in Clark and McCracken (2007), algebra along the lines of McCracken (2000) proves the consistency of $\hat{\Gamma}_{dd}(j)$ and $\hat{\Gamma}_{dh}(j)$. (ii) Given part (i)–and especially that $r_T = o_p(P^{-m})$–the proof is identical to that for Theorem 2.3.2 in McCracken (2000).

Table 1. Non-Nested Model Size and Power Results

test	P = 20	40	80	160	P = 20	40	80	160
	horizon = 1				horizon = 4			
	size: DGP 1, news							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.22	.12	.08	.06
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.04
	size: DGP 2, news							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.22	.12	.08	.06
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.04
	size: DGP 3, news							
MSE-$t(S_{dd})$	.10	.08	.07	.06	.28	.16	.11	.08
MSE-$t(\Omega)$	.05	.04	.04	.05	.14	.09	.06	.06
	size: DGP 1, noise							
MSE-$t(S_{dd})$	.11	.11	.11	.12	.23	.14	.10	.08
MSE-$t(\Omega)$	.06	.04	.04	.04	.09	.05	.03	.03
	size: DGP 2, noise							
MSE-$t(S_{dd})$	.09	.07	.06	.05	.21	.12	.07	.05
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.03
	size: DGP 3, noise							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.27	.15	.09	.07
MSE-$t(\Omega)$	.05	.04	.04	.04	.14	.08	.05	.05
	power: DGP 1, news							
MSE-$t(S_{dd})$	.70	.94	1.00	1.00	.26	.24	.32	.53
MSE-$t(\Omega)$	.70	.94	1.00	1.00	.20	.19	.28	.52
	power: DGP 2, news							
MSE-$t(S_{dd})$	.48	.75	.95	1.00	.26	.23	.30	.50
MSE-$t(\Omega)$	.48	.74	.96	1.00	.20	.18	.26	.48
	power: DGP 3, news							
MSE-$t(S_{dd})$	.34	.52	.80	.97	.33	.29	.38	.60
MSE-$t(\Omega)$	.32	.52	.80	.97	.25	.27	.38	.62
	power: DGP 1, noise							
MSE-$t(S_{dd})$	.12	.09	.10	.11	.22	.14	.10	.08
MSE-$t(\Omega)$	.09	.06	.06	.07	.13	.07	.05	.04
	power: DGP 2, noise							
MSE-$t(S_{dd})$	.22	.33	.54	.83	.23	.16	.16	.22
MSE-$t(\Omega)$	.22	.33	.55	.84	.17	.12	.14	.22
	power: DGP 3, noise							
MSE-$t(S_{dd})$	.11	.10	.11	.14	.24	.15	.11	.10
MSE-$t(\Omega)$	.10	.09	.11	.14	.17	.12	.10	.11

Notes:
1. The DGPs and forecasting equations are given in section 4.1. In the size experiments, the DGP coefficient β is set to 0. In the power experiments, the DGP coefficient β is set to 0.6 for DGP 1 and 2 and 0.75 for DGP 3.
2. R, the size of the sample used to generate the first forecast, is 80. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000. The nominal size is 5%.
3. MSE-$t(S_{dd})$ refers to an unadjusted t-test for equal MSE, using the conventional variance $\hat{S}_{dd}$. MSE-$t(\Omega)$ refers to an adjusted t-test for equal MSE, using the variance $\hat{\Omega} = \hat{S}_{dd} + 2\hat{\Pi}(\hat{F}\hat{B}\hat{S}_{dh} + \hat{F}\hat{B}\hat{S}_{hh}\hat{B}\hat{F}')$. All test statistics are compared against standard normal critical values.

Table 2. Nested Model Size and Power Results

test	c.v.	P = 20	40	80	160	P = 20	40	80	160
		horizon = 1				horizon = 4			
		size: DGP 1							
MSE-F	CM	.05	.11	.18	.26	.06	.11	.17	.25
MSE-$t(S_{dd})$	CM	.21	.23	.27	.32	.25	.22	.23	.26
MSE-$t(S_{dd})$	N	.10	.12	.17	.24	.16	.13	.14	.17
MSE-$t(\Omega)$	N	.10	.09	.09	.08	.15	.14	.13	.13
CCS	χ^2	.07	.05	.04	.04	.21	.12	.09	.07
		size: DGP 2							
MSE-F	CM	.04	.05	.06	.07	.04	.05	.06	.05
MSE-$t(S_{dd})$	CM	.12	.08	.06	.05	.19	.11	.06	.03
MSE-$t(S_{dd})$	N	.04	.02	.02	.01	.10	.05	.02	.01
MSE-$t(\Omega)$	N	.10	.10	.08	.07	.12	.11	.08	.06
CCS	χ^2	.08	.06	.05	.05	.23	.13	.09	.08
		size: DGP 3							
MSE-F	CM	.05	.07	.09	.10	.05	.07	.08	.09
MSE-$t(S_{dd})$	CM	.13	.10	.08	.07	.21	.14	.10	.06
MSE-$t(S_{dd})$	N	.04	.03	.02	.02	.12	.06	.03	.02
MSE-$t(\Omega)$	N	.12	.11	.10	.09	.14	.14	.11	.09
CCS	χ^2	.09	.07	.07	.06	.22	.14	.09	.08
		power: DGP 1							
MSE-F	CM	.39	.64	.87	.98	.25	.45	.70	.91
MSE-$t(S_{dd})$	CM	.52	.66	.83	.96	.48	.53	.65	.83
MSE-$t(S_{dd})$	N	.28	.41	.63	.88	.33	.34	.44	.65
MSE-$t(\Omega)$	N	.63	.72	.83	.93	.50	.56	.67	.81
CCS	χ^2	.07	.06	.06	.08	.20	.13	.12	.15
		power: DGP 2							
MSE-F	CM	.23	.31	.39	.48	.10	.16	.22	.28
MSE-$t(S_{dd})$	CM	.20	.20	.21	.23	.24	.19	.16	.15
MSE-$t(S_{dd})$	N	.08	.07	.07	.09	.14	.08	.06	.05
MSE-$t(\Omega)$	N	.38	.38	.39	.40	.31	.29	.28	.26
CCS	χ^2	.09	.10	.14	.24	.22	.15	.17	.26
		power: DGP 3							
MSE-F	CM	.26	.29	.33	.34	.15	.21	.29	.36
MSE-$t(S_{dd})$	CM	.18	.16	.14	.13	.27	.22	.21	.20
MSE-$t(S_{dd})$	N	.07	.06	.05	.04	.16	.11	.08	.08
MSE-$t(\Omega)$	N	.36	.32	.28	.23	.34	.34	.33	.32
CCS	χ^2	.11	.14	.26	.48	.20	.14	.16	.25

Notes:
1. The DGPs and forecasting equations are given in section 4.2. The DGP coefficient β_{22} is set to 0 in size experiments and 0.3 in power experiments.
2. R, the size of the sample used to generate the first forecast, is 80. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000. The nominal size is 5%.
3. The tests are defined as follows: MSE-F = F-test for equal MSE; MSE-$t(S_{dd})$ = t-test for equal MSE, using a variance $\hat{S}_{dd}$; MSE-$t(\Omega)$ = t-test for equal MSE using a variance $\hat{\Omega}$ =2 $\hat{\Pi}\hat{F}(-J\hat{B}_1J' + \hat{B}_2)\hat{S}_{hh}(-J\hat{B}_1J' + \hat{B}_2)\hat{F}'$; and CCS = the Chao, Corradi, and Swanson (2001) test. A 'CM' in column two means the critical value is obtained from the simulation method of Clark and McCracken (2005); 'N' means the critical value is taken from the standard normal distribution; and χ^2 indicates critical values taken from that distribution.

Table 3. Results for Inflation Forecast Application

sample	MSE_1	MSE_2	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	MSE-$t(S_{dd})$	MSE-$t(\Omega)$	MSE-F	CCS
non-nested models								
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.268	2.022	.269	.364	.914	.675	NA	NA
1970-1984	4.060	3.844	.598	.781	.361	.277	NA	NA
1985-2003	.924	.655	.152	.232	1.765^a	1.159	NA	NA
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.262	1.937	.272	.393	1.196	.828	NA	NA
1970-1984	4.144	3.715	.605	.840	.709	.511	NA	NA
1985-2003	.851	.604	.143	.222	1.728^a	1.114	NA	NA
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.468	1.997	.280	.451	1.683^a	1.044	NA	NA
1970-1984	4.599	3.841	.608	.966	1.246	.785	NA	NA
1985-2003	.870	.614	.127	.228	2.019^b	1.121	NA	NA
nested models								
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.676	2.022	.423	.140	1.545^c	4.689^c	43.038^c	4.189
1970-1984	5.340	3.844	.855	.490	1.749^c	3.055^c	22.194^c	5.372
1985-2003	.678	.655	.131	.042	.171	.533	2.605	5.195
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.574	1.937	.414	.131	1.538^c	4.867^c	43.730^c	2.732
1970-1984	5.148	3.715	.839	.430	1.710^c	3.332^c	22.001^c	4.081
1985-2003	.644	.604	.143	.024	.275	1.665^b	4.954^a	2.518
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.480	1.997	.386	.110	1.251^c	4.398^c	32.169^c	2.057
1970-1984	5.046	3.841	.788	.423	1.528^c	2.847^c	17.879^c	5.480
1985-2003	.556	.614	.131	.054	-.444	-1.083	-7.200	2.339

Notes:

1. The table compares the accuracy of real-time forecasts of the one-year-ahead change in GDP inflation. MSE_i denotes the mean square forecast error from model i. The forecasts in the non-nested comparison are generated from equation (22), with model 1 using x_t = the output gap (computed with the HP filter) and model 2 using x_t = four-quarter GDP growth. The forecasts in the nested comparison are generated from equations (23) (model 1) and (22) (model 2). The models are estimated recursively, with the sample beginning in 1961:1+τ-1.

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 first available estimate of inflation; the next the second 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 (2005), and the MSE-$t(\Omega)$ and CCS statistics are compared against, respectively, standard normal and χ^2 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}.