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

April 2007
(preliminary)

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 regression models. In contrast to earlier work — including West (1996) and Clark and McCracken (2001) — 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: Prediction, real-time data, 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 suggestions from Borağan Aruoba and helpful comments from seminar participants at Oregon, SUNY-Albany, and UBC. 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

In practice it is often the case that more than one model is available for forecasting. This arises easily in economics and finance where different theories regarding the behavior of economic agents can imply that different variables and models should have predictive content. With forecasts from these various models in hand, one might reasonably ask whether one of the models forecasts more accurately than the other and if the difference is statistically significant.

As such, 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. Various methods have been developed to test whether any gains from the new model are statistically significant. 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. 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; the tests continue to be asymptotically standard normal. For nested models, Clark and McCracken (2001, 2006), McCracken (2006), Chao, Corradi and Swanson (2001), and Corradi and Swanson (2002, 2005) derive asymptotics for a collection of tests designed to determine whether a nested model forecasts as accurately or encompasses a larger, nesting, model. In most cases, nested comparisons imply asymptotic distributions that are not asymptotically standard normal. However, all of these studies rely on on asymptotics that treat the estimation and forecasting samples as limiting to infinity (covering both recursive and rolling forecasting schemes). Under an alternative asymptotic approximation that treats the estimation sample as fixed (as in a rolling forecasting scheme) rather than limiting to infinity, Giacomini and White (2006) obtain asymptotic normality for a test of equal predictive ability.

While this literature is rich with results and continues to grow (see such recent contributions as Armah and Swanson (2006) and Anatolyev (2007)), one issue that is uniformly overlooked is the real-time nature of the data. Specifically, the literature ignores the possibility that at any given forecast origin the most recent data is subject to revision. At first, a short-lived revision process may seem unlikely to have much of an effect on the asymptotic distribution of a test statistic. As an example, consider the standard F-test for predictive ability constructed using observations $t = 1, ...T$. If the final observation is subject to revision, so long as the revision is finite the asymptotic distribution of the F-test, taken as $T \rightarrow \infty$, will be unaffected because, under reasonable assumptions, a single observation will almost surely have no influence on the parameter estimates and subsequent test statistic.

Now consider the case in which an out-of-sample test of predictive ability is being constructed. The test statistic is functionally very different from an in-sample one and in a fashion that makes it particularly susceptible to changes in the correlation structure of the data as the revision process unfolds. This occurs for three reasons: (i) while parameter estimates are typically functions of only a small number of observations that remain subject to revision, out-of-sample statistics are themselves functions of a sequence of these 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 it is the case, as noted in Aruoba (2006), that data subject to revision possess a different mean and covariance structure than final revised data, it is not surprising that tests of predictive ability using real-time data may have a different asymptotic distribution than tests constructed using data that is never revised.

Accordingly, in this paper we provide analytical, Monte Carlo and empirical evidence on pairwise tests of equal out-of-sample predictive ability for models estimated — and forecasts evaluated — using real-time data. We consider comparisons whereby the models are non-nested or nested, as well as a design we refer to as reverse-overlapping. In each case we restrict attention to linear direct multi-step (DMS) models evaluated under quadratic loss but do not require that the models be correctly specified; model residuals and forecast errors are allowed to be conditionally heteroskedastic and serially correlated of an order greater than the forecast horizon. In some cases, we permit the revision process to consist of both "news" and "noise" as defined in Mankiw, Runkle and Shapiro (1984) and applied more recently by Aruoba (2006). In general, though, we emphasize the role of noisy revisions.

Our results indicate substantial differences in the asymptotic behavior of tests of equal predictive ability, relative to those found in the existing literature, when data is subject to revisions. For example, when constructing tests of equal predictive ability between non-nested models, West (1996) notes 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 in which the revision process consists only of news. When even some noise is present, parameter estimation error contributes to the asymptotic variance of the test statistic and cannot be ignored when conducting inference.

As another example, when constructing tests of equal predictive ability between nested models, Clark and McCracken (2001, 2005) and McCracken (2006) note that standard test statistics used to evaluate predictive ability 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 fail not only to be asymptotically normal, but in fact diverge with probability one under the null hypothesis. To avoid this, we introduce a variant of the standard test statistic that is asymptotically normal despite being a comparison between two, recursively estimated, nested models.

In the case of predictable revisions, we also consider a new situation we refer to as reverse-overlapping. The term "overlapping" comes from Vuong (1989) and describes a situation where the null hypothesis of equal in-sample predictive ability between two ostensibly non-nested models can arise two ways — each leading to a distinct asymptotic distribution. In the first, the two models are non-nested with a non-degenerate (in-sample) loss differential and asymptotic normality is obtained for the likelihood ratio. In the second, the two models collapse onto a single model that is nested within each model and the likelihood ratio is asymptotically mixed chi-square. In our case, the *reverse* is true: an ostensibly *nested* pair of models can satisfy the null two ways (described below), each leading to a distinct asymptotic distribution. While each is asymptotically normal, the appropriate asymptotic variance can be very different. An example will be provided in Section 3.3.

Not surprisingly, as with all theoretical results, our conclusions rely upon assumptions made on the observables. What makes our problem specifically troublesome 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 dependent or predictor variable. To keep our analytics as transparent as possible, while still remaining relevant for application, we assume that for each variable the revision process continues sequentially for a finite $0 \leq r << R$ periods. While these revisions are assumed to be covariance stationary, only limited assumptions are made directly on the observables across revisions for a fixed historical date. Importantly, we also abstract from other forms of revisions including benchmark revisions. We leave these important issues to subsequent research.

While our results are related to the existing literature on tests of out-of-sample predictability, our results also relate back to a literature on forecasting in the presence of data revisions including Howrey (1978), Swanson (1996) and Robertson and Tallman (1998). Notably, our results bear some resemblance to those in Koenig, Dolmas and Piger (2003). They, too, note that the observables likely have different statistical properties depending upon where the observables are in the revision process. They suggest that one can improve forecast accuracy by using the various vintages of data as they would have been observed in real-time to construct forecasts rather than only using those observables that exist in the most recent vintage. Their results differ from ours in that they are interested in forecast accuracy while we are interested in out-of-sample inference but the main issue remains the same: ignoring the data revision process can lead to undesired outcomes — either less accurate forecasts or, in our case, asymptotically invalid inference.

The remainder of 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 of the asymptotic results 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 determine whether measures of output have predictive content for 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 being 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. 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 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 or reverse-overlapping 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 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).

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, we keep the vintage horizon $r' = t' - t - \tau$ fixed. At the initial forecast origin $t = R$, the present data 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 span 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 $\overline{T} = T + \tau + r'$. Note that the final $\tau + r'$ vintages are used exclusively for evaluation.

Forecasts of $y_{t+\tau}(t')$, $t = R, \ldots, T$, are generated using the 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 $s = 1, ..., t-\tau$. Under the null hypothesis of equal forecast accuracy between nested models, model 2 nests model 1 for all t such that model 2 includes $\dim(x_{22,s}(t)) = k_{22}$ excess parameters. Then $\beta^*_2 = (\beta^{*\prime}_1, 0')'$, 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'. Because of this degeneracy, the hypothesis of equal population predictive ability is trivially true since $Eu^2_{1,t+\tau}(t') = Eu^2_{2,t+\tau}(t') \equiv Eu^2_{t+\tau}(t')$ for all t and t'.

Under the null hypothesis of equal forecast accuracy between non-nested or reverse-overlapping models, there are no explicit restrictions on the model parameters. We only require that, when evaluated at the population value of the pseudo-true parameters associated with the models, the squared forecast errors have a common mean and hence (with a covariance stationarity assumption made later) $E(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t')) = 0$ for all t.

Both model 1's and model 2's forecasts are generated recursively using OLS-estimated parameters. Under this approach 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 $\hat{u}_{2,t+\tau}(t') = y_{t+\tau}(t') - x'_{2,t}(t)\hat{\beta}_{2,t}$, respectively.

Finally, the asymptotic results below use the following additional notation. Let $h_{i,t+\tau}(t') = (y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$, $h_{i,s+\tau} = (y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$, $H_i(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{i,s+\tau}$, $B_i = (Ex_{i,s}x'_{i,s})^{-1}$ and $d_{t+\tau}(t') = u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t')$. Throughout, when the models are non-nested or reverse-overlapping 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 $h_{s+\tau} = h_{2,s+\tau}$, $h_{t+\tau}(t') = h_{2,t+\tau}(t')$ and $U_{t+\tau} = [h'_{t+\tau}(t') - Eh'_{t+\tau}(t'), h'_{t+\tau}, x'_t - Ex'_t]'$.[1] In either case let $H(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$. 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 loss differential $d_{t+\tau}(t')$ defined more precisely in Section 3.

Given the definitions and forecasting scheme described above, the following assumptions are used to derive the limiting distributions in Theorems 1-4. The assumptions are intended to be only sufficient, not necessary and sufficient.

(A1) The parameter estimates $\hat{\beta}_{i,t}$, $i = 1, 2$, $t = R, ..., T$, are estimated using OLS for each vintage in succession and hence satisfy $\hat{\beta}_{i,t} = \arg\min_\beta \sum_{s=1}^{t-\tau}(y_{s+\tau}(t) - x'_{i,s}(t)\beta_i)^2$.[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 for each integer $0 \leq j$, $(y_t(t+j), x'_t(t+j))'$ is uniformly L^{2n} bounded, (e) $U_{t+\tau}$ is strong mixing with coefficients of size $-2n/(n-1)$, (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 $\int_{-\infty}^{\infty}|K(x)|dx$, (b) For some bandwidth M and constant $i \in (0, 0.5)$, $M = O(P^i)$.

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

(A4′) $\lim_{R,P\to\infty} P/R = 0$.

The assumptions provided here are closely related to those in West (1996) and Clark and McCracken (2005). We restrict attention to forecasts generated using parameters estimated by OLS (Assumption 1) and we do not allow for processes with either unit roots

[1] When the models are nested, $d_{t+\tau}(t') = 0$ for all t and t'.

[2] That is, at each forecast origin t, we use the last vintage of data available at period t to estimate the model by OLS. As forecasting moves forward in time, we use ever-newer vintages of data, and a time sample of increasing length.

or time trends (Assumption 2). When long-run variances are estimated, standard kernel estimators are used (Assumption 3). 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$'$).

Although our assumptions are restrictive in some ways — notably the comparison of linear models — in other ways they are fairly general. We allow for conditional heteroskedasticity and serial correlation in the levels and squares of the forecast errors. Nevertheless, our assumptions remain strong enough for us to use Wooldridge and White's (1998) theoretical results on CLTs.

3 Tests and Asymptotic Distributions

In this section we provide asymptotics for tests of equal forecast accuracy for non-nested, nested and reverse-overlapping comparisons. For the comparison of non-nested models we allow data revisions to consist of both news and noise. For reverse-overlapping model comparisons, noisy revisions are the only relevant form of revision. In the nested case, for tractability we allow the data revisions to consist only of noise.[3]

In each case, we begin by presenting asymptotically valid expansions of the sample average of the loss differentials associated with models 1 and 2, $(P-\tau+1)^{-1}\sum_{t=R}^{T}(\hat{u}_{1,t+\tau}^{2}(t')-\hat{u}_{1,t+\tau}^{2}(t'))$. We present these expansions in order to make clear exactly how and when data revisions affect the asymptotic distribution of the tests of equal forecast accuracy already existing in the literature. Building upon these expansions, we then provide theorems that characterize the asymptotic distributions of certain test statistics emphasizing how asymptotically valid inference can be conducted in the presence of data revisions.

3.1 Non-nested comparisons

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal MSE based upon the sequence of loss differentials $\hat{d}_{t+\tau}(t')=\hat{u}_{1,t+\tau}^{2}(t')-\hat{u}_{2,t+\tau}^{2}(t')$. If we define $MSE_i=(P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{u}_{i,t+\tau}^{2}(t')$ $(i=1,2)$, $\bar{d}=(P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{d}_{t+\tau}(t')=MSE_1-MSE_2$, $\hat{\Gamma}_{dd}(j)=(P-\tau+1)^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t')-\bar{d})(\hat{d}_{t+\tau-j}(t'-j)-\bar{d})$, $\hat{\Gamma}_{dd}(-j)=$

[3] Working with revisions consisting of news is feasible but an order of magnitude more complex than for the non-nested case. We will return to this issue in a subsequent draft of the paper.

$\hat{\Gamma}_{dd}(j)$, 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 and hence inference can be conducted using the relevant tables.

West (1996) however, notes that this outcome depends upon whether or not the forecast errors depend upon estimated parameters. If they do, then the statistic remains asymptotically normal but may have an asymptotic variance that reflects not only the long-run variance of the loss differential $\lim_{R,P\to\infty}\ var(P^{1/2}\bar{d}) = S_{dd}$ but also additional variance and covariance terms that arise due to parameter estimation error. 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_{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 missized 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 $F = 0$. This case arises naturally in the present context because F is equal to 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.

In the presence of data revisions, it's this last part that draws attention — that the observables used to construct and evaluate the forecast errors are covariance stationary. For example, in the absence of data revisions, $y_s(t) = y_s(t')$ and $x_s(t) = x_s(t')$ for all t, t'. Hence at the population level, the residuals $y_{s+\tau} - x'_{i,s}\beta_i^*$, $s = 1, ..., t-\tau$, and forecast errors $y_{t+\tau} - x'_{i,t}\beta_i^*$, $t = R, ..., T$, have the same covariance structure. This implies that when the in-sample moment condition $E(y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s} = 0$ is satisfied it must also be the case that the out-of-sample moment condition $E(y_{t+\tau} - x'_{i,t}\beta_i^*)x_{i,t} = 0$ is satisfied. But when there are data revisions, $y_{s+\tau} - x'_{i,s}\beta_i^*$ and $y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*$ 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 the moment $E(y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$ equals zero.

If we keep track of this distinction and use algebra along the lines of West (1996), we obtain the following expansion.

Lemma 1: 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. Conceptually, though, it differs in two important ways. First, the analytics are derived allowing for data revisions at the end of each sequential vintage of data. Second, the term F is defined as $2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$, thus emphasizing the distinction between the population in-sample residuals and the population out-of-sample forecast errors. Since the asymptotic expansion is notationally identical to that in West (1996) it's not surprising that 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. $P^{1/2}\bar{d} \to^d N(0,\Omega)$ 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 $P-\tau+1$ is small relative to the number of in-sample observations from the initial forecast origin R, 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. This case is identical to that in West (1996).

Another special case arises when the out-of sample moment condition $F = 2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$ 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 (2001))

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, 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 given in Theorem 1. We return to consistent estimation of Ω in Section 3.4.

3.2 Nested comparisons

In the context of nested models, Clark and McCracken (2005) and McCracken (2006) also propose tests for equal MSE based upon the sequence of loss differentials. Specifically, they consider the MSE-t statistic discussed in (1) but applied to nested models and another that can be constructed analogously to an in-sample F-test but using out-of-sample forecast errors, given by

$$\text{MSE-}F = (P - \tau + 1) \times \frac{\text{MSE}_1 - \text{MSE}_2}{\text{MSE}_2} = (P - \tau + 1) \times \frac{\bar{d}}{\text{MSE}_2}. \tag{2}$$

In both cases, the tests have limiting distributions that are non–standard when the forecasts are nested under the null. Specifically, McCracken (2006) show thats, 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}, but not any unknown nuisance parameters. 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. Accordingly, for this situation, a bootstrap procedure is recommended. 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 — much more so than in the non-nested case. As was the case for non-nested models, the crux of the problem is that when there are 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, in particular, $E(y_{s+\tau} - x'_{2,s}\beta^*_2)x_{2,s}$

equaling zero need not imply anything about whether or not the moment $F = 2E(y_{t+\tau}(t') - x'_{2,t}(t)\beta_2^*)x_{2,t}(t)$ equals zero. If we keep track of this distinction and use algebra along the lines of West (1996), we obtain the following expansion.

Lemma 2: Let Assumptions 1 and 2 hold and let $F \neq 0$. (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 $(P^{-1/2}\sum_{t=R}^{T} u_{1,t+\tau}^2(t') - u_{2,t+\tau}^2(t'))$ 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 (2006) for nested models. The key difference is that the Lemma 2 expansion is of order $P^{1/2}$, rather than the order P in Clark and McCracken (2005) and McCracken (2006) and as one would typically expect from a comparison of nested models using a statistic like an F-stat. Not surprisingly, this change in order implies very different asymptotic behavior of out-of-sample averages of loss differentials from nested models.

Theorem 2: Let Assumptions 1 and 2 hold and let $F \neq 0$. (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 normal under the null hypothesis. This is in sharp contrast to the results in Clark and McCracken (2005) and McCracken (2006), in which the tests generally have non-standard limiting distributions. This finding has a number of important implications, listed below.

1. The MSE-F statistic diverges with probability 1 under the null hypothesis. To see this note that Theorem 2 implies that the numerator of MSE-F is $P^{1/2}(P^{1/2}\bar{d})$. So long as the probability limit of MSE_2 is finite we know that the MSE-F is $O_p(P^{1/2})$ and hence the asymptotic size of the test (one-sided to the right) is 50%.

2. The MSE-t test (1) also 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 (2006), the denominator of the MSE-t 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 also has an asymptotic size of 50%.

3. The standard forms of the MSE-F and MSE-t tests have power against the null of Granger non-causality and the news vs. noise hypothesis that the in-sample F-test does not have.

4. Out-of-sample inference for nested comparisons can be conducted without the strong auxiliary assumptions made in Clark and McCracken (2005) and McCracken (2006) regarding the correct specification of the models.[4]

5. 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 3.4.

However, even with predictable revisions (that make F non-zero), it is possible that the asymptotic distributions of the MSE-t and MSE-F tests can differ from the results given in Theorem 2. In some cases, even with $F \neq 0$, the variance Ω may be zero, due to a singularity in the middle term of the quadratic form that determines Ω (specifically, $(-JB_1J' + B_2)$). Cancellation among terms in $(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)$ may make Ω singular. This cancellation seems more likely to occur with one-step forecasts and conditionally homoskedastic residuals (which reduce S_{hh} to $\sigma^2 B_2^{-1}$), although it may occur even without these restrictions. As a simple example, suppose a DGP that relates y to a variable x with mean $\mu \neq 0$, with both y and x having variances of 1. Let y be subject to one revision (provided one period after the publication of the initial estimate), with revisions having a mean of $\kappa \neq 0$. Suppose the null forecasting model includes just a constant; the alternative includes a constant and x. In this case, the F vector is $-2\kappa(1, \mu) \neq 0$. However, working through the algebra shows that $\Omega = 0$. In such situations, the numerator and denominator of a t-test for equal MSE will both be converging to zero. If the convergence rates are the same, the test may have a non–degenerate distribution, but that distribution will differ from those described above. Nonetheless, it is possible that, in finite samples, the distributions described above may reasonably approximate the actual distribution. Using this simple DGP, we examine this possibility in the Monte Carlo analysis.

[4] In previous work we have required that serial correlation in the residuals and forecast errors were of finite order. In most instances we treated τ-step ahead errors as forming an MA$(\tau - 1)$ process.

3.3 Reverse overlapping comparisons

In the preceding discussion, the null hypothesis of equal forecast accuracy between two nested models was imposed by maintaining that the additional predictors, $x_{22,t}$, associated with the unrestricted model held no predictive content for $y_{t+\tau}$ and hence the associated regression parameters β_{22} were zero. While not immediately obvious, the null hypothesis of equal forecast accuracy can hold even when β_{22} is not zero — but only when it is also the case that the data used for evaluating the forecast is subject to revision and has the right covariance structure.

To see this consider the following simple example. Suppose that the dependent variable (which is subject to revision) is determined by the covariance stationary, simple, linear regression $y_{t+1} = \beta_0 + \beta_{22}x_t + \varepsilon_{t+1}$, with scalar stochastic regressor x_t (that is not revised) and white noise error ε_t. Let model 1 be the trivial constant mean model consisting of just an intercept and let model 2 be the correctly specified model consisting of both an intercept and x_t. If we estimate each by OLS, the associated squared forecast errors, evaluated at the probability limits of the parameter estimates, are $u^2_{1,t+1}(t') = (y_{t+1}(t') - Ey_{t+1})^2$ and $u^2_{2,t+1}(t') = (y_{t+1}(t') - Ey_{t+1} - (x_t - Ex_t)\beta_{22})^2$, where $\beta_{22} = cov(x_t, y_{t+1})/var(x_t)$.

For equal forecast accuracy the expected difference of these squared forecast errors should be zero. Taking this expectation we obtain $E[u^2_{1,t+1}(t') - u^2_{2,t+1}(t')] = -2E[(y_{t+1} - y_{t+1}(t'))(x_t - Ex_t)]\beta_{22} + E(x_t - Ex_t)^2\beta^2_{22}$. In the previous section, we obtained equal forecast accuracy because β_{22} was restricted to zero under the null. Closer inspection, however, reveals that the difference can be zero even when β_{22} is not, so long as the revision $y_{t+1} - y_{t+1}(t')$ is not zero as well. Substituting in the definition of β_{22} and rearranging terms we find that $E[u^2_{1,t+1}(t') - u^2_{2,t+1}(t')]$ can also be zero if $2cov(y_{t+1} - y_{t+1}(t'), x_t) = cov(x_t, y_{t+1})$. For this case to apply, the revisions $y_{t+1} - y_{t+1}(t')$ must have just the right covariance with the predictors x_t. Equivalently, the two models will have equal predictive ability so long as $\beta_{22} = 2cov(y_{t+1} - y_{t+1}(t'))/var(x_t)$ — that is, if the regression coefficient happens to be twice the value of the slope coefficient associated with the projection of the revision $y_{t+1} - y_{t+1}(t')$ on the predictand x_t.

When this situation arises we refer to two models as being reverse-overlapping. The term "overlapping" comes from Vuong (1989) and describes a situation in which the null hypothesis of equal in-sample predictive ability between two ostensibly non-nested models can arise two ways, each leading to a distinct asymptotic distribution. In our case, the "reverse" is true: an ostensibly *nested* pair of models can satisfy the null two ways, each leading to a distinct asymptotic distribution. The first is precisely that from our out-of-sample theory in the previous nested section. The latter, as we will see below, is related to our out-of-sample theory from the previous non-nested section.

Prior to presenting the result it is helpful to revisit some notation. Recall from Section 2 that, when we work with reverse-overlapping models, we define $h_{t+\tau} = (h'_{1,t+\tau}, h'_{2,t+\tau})'$ as we did in the non-nested case, rather than $h_{t+\tau} = h_{2,t+\tau}$ as we did in the nested case. Similarly, we define F as $2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$ rather than $2Eu_{2,t+\tau}(t')x_{2,t}(t)$ as we did for the nested case. With these changes in hand, if we again let B denote the block diagonal matrix formed by B_1 and B_2, we obtain the following expansion.

Lemma 3: 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 3 matches that of Lemma 1 for the case of non-nested models rather than that of Lemma 2 for the nested case. The reason is that the interaction of the additional predictive content in $x_{22,t}$ with this special form of data revision (i) induces the initial term, $P^{-1/2}\sum_{t=R}^{T}(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))$, to be non-zero and (ii) prevents $h_{1,t+\tau}(t')$ and $h_{1,t+\tau}$ from being numerically equivalent to $Jh_{1,t+\tau}(t')$ and $Jh_{2,t+\tau}$ respectively so that F and $h_{t+\tau}$ need to be redefined. Again, since the asymptotic expansion is identical to that in West (1996), it's not surprising that the asymptotic distribution of the scaled average of the loss differentials remains (notationally) the same.

Theorem 3: Let Assumptions 1, 2 and 4 or 4$'$ hold. $P^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$ where $\Omega = S_{dd} + 2(1 - \pi^{-1}\ln(1+\pi))(FBS_{dh} + FBS_{hh}BF')$.

Because of the similarity between Theorem 1 and Theorem 3, some of the methods that apply to constructing an asymptotically valid test statistic for the non-nested case remain applicable for the reverse-overlapping case. When $\pi = 0$, $2(1 - \pi^{-1}\ln(1+\pi))$ equals zero, and hence the effects of parameter estimation error are asymptotically negligible. In contrast, though, by the very nature of the reverse-overlapping models, it is very unlikely that it will be the case that $F = 0$. To see this, note that, using our earlier example as a foil, since β_{22} is not zero neither is $cov(x_t, y_{t+\tau})$. Since reverse overlapping implies a non-degenerate relationship between $cov(x_t, y_{t+\tau})$ and $cov(x_t, y_{t+\tau} - y_{t+\tau}(t'))$, it must also be the case that $cov(x_t, y_{t+\tau} - y_{t+\tau}(t'))$ is non-zero. But for our simple example, this in turn implies that F cannot be zero.

For reverse-overlapping comparisons it is then again the case that conducting asymptotically valid inference will require consistent estimation of the appropriate standard errors from Theorem 3. We show how to do so in the following section.

3.4 Estimating the standard errors

For each of the model comparisons we obtain results suggesting that $P^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$ (or $R^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$) will be asymptotically standard normal and hence the corresponding tables can be used to conduct inference so long as consistent estimates of the relevant standard errors can be constructed. In this section we provide details on methods for constructing asymptotically valid estimates of the standard errors associated with each of the non-nested, nested and reverse-overlapping cases.

In each case, 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 F and B we use the obvious sample analogs. 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 and reverse-overlapping 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 upon standard kernel-based estimators akin to those used in West (1996), West and McCracken (1998) and McCracken (2000). To be more precise, we use kernel-weighted estimates of $\Gamma_{dd}(j) = E(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))(u^2_{1,t+\tau-j}(t'-j) - u^2_{2,t+\tau-j}(t'-j))$, $\Gamma_{dh}(j) = E(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))h'_{t+\tau-j}$ 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 and reverse-overlapping 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 $\widehat{h}_{s+\tau} = (y_{s+\tau} - x'_{2,s}\hat{\beta}_{2,T})x_{2,s}$, $s = 1, ..., T$.

With these sequences of forecast errors and OLS orthogonality conditions in hand, let $\hat{\Gamma}_{dd}(j) = P^{-1}\sum_{t=R+j}^{T}(\hat{u}^2_{1,t+\tau}(t') - \hat{u}^2_{2,t+\tau}(t'))(\hat{u}^2_{1,t+\tau-j}(t'-j) - \hat{u}^2_{2,t+\tau-j}(t'-j))$, $\hat{\Gamma}_{hh}(j) = T^{-1}\sum_{s=1+j}^{T}\widehat{h}_{s+\tau}\widehat{h}'_{s+\tau-j}$ and $\hat{\Gamma}_{dh}(j) = P^{-1}\sum_{t=R+j}^{T}(\hat{u}^2_{1,t+\tau}(t') - \hat{u}^2_{2,t+\tau}(t'))\widehat{h}'_{t+\tau-j}$, with $\hat{\Gamma}_{dd}(j) = \hat{\Gamma}_{dd}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}'_{hh}(-j)$ and $\hat{\Gamma}_{dh}(j) = \hat{\Gamma}'_{dh}(-j)$. Let $K(.)$ define an appropriate kernel function and M a bandwidth. We then estimate the long-run variances and covariances as $\hat{S}_{dd} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dd}(j)$, $\hat{S}_{hh} = \sum_{j=-T+1}^{T-1}K(j/M)\hat{\Gamma}_{hh}(j)$, and $\hat{S}_{dh} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dh}(j)$. The following theorem shows that the relevant pieces are consistent for their population counterparts.

Theorem 4: 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 1-3, Theorem 4 and Slutsky's Theorem imply that $P^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$ (or $R^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$) is asymptotically standard normal and hence asymptotically valid inference can be conducted using the appropriate tables. Monte Carlo evidence on the finite sample performance of these estimators is given in Section 4.

Of course valid inference requires using the individual components appropriately when constructing $\widehat{\Omega}$. 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 that there is a noise component to the data revisions (with no noise, either estimate is asymptotically appropriate; with noise, only the former is appropropriate). For reverse-overlapping comparisons the former is the only relevant choice because predictable revisions are a necessary condition for the comparison to exist. For nested comparisons, 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.

Interestingly, when making an ostensibly nested comparison, if one is unsure of whether or not the reverse-overlapping case might apply, one can use the reverse-overlapping variant of $\widehat{\Omega}$ as an asymptotically valid estimate of both $S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$ and $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$. To see this, note that when the models are nested, S_{dd} and S_{dh} are both zero and furthermore, straightforward algebra reveals that the reverse-overlapping definition of $FBS_{hh}BF'$ collapses to the nested definition of $F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$. However, in the Monte Carlo experiments described in the next section, it was consistently the case that using the standard error estimate based on the overlapping approach yielded a rejection rate of roughly zero. As a result, we don't include this approach in the nested model simulation results below.

4 Monte Carlo Evidence

We use simulations of simple DGPs to evaluate the finite sample properties of the above tests for equal forecast MSE in the presence of data revisions that exhibit either news or noise. As a baseline we also consider results in the absence of revisions. In simulations of non–nested models, we focus on t-tests for equal MSE, one computed without regard to the impact of data revisions (using just S_{dd}) and another adjusting for the impact of data revisions (using $S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF'))$. In simulations of the nested model case, we evaluate size and power of tests of equal forecast accuracy using critical values based on previous work (McCracken, 2006) that ignores the revision process as well as those that use standard normal critical values that are applicable in the presence of predictable revisions. For the nested experiments we focus on $\pi > 0$ asymptotics and consider four tests: MSE-F compared against critical values from McCracken (2006), MSE-t using $\Omega = S_{dd}$ and critical values from McCracken (2006), MSE-t using $\Omega = 2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ and standard normal critical values, and MSE-t using $\Omega = S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$ and standard normal critical values. We also report results for three test statistics based on $\pi = 0$ asymptotics: MSE-F against $\pi = 0$ critical values from McCracken (2006), MSE-t using S_{dd} and standard normal critical values, and MSE-t using $\Omega = F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ and standard normal critical values.

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, first for the non–nested case and then the nested case. In this draft, we consider just one–step ahead forecasts (so $\tau = 1$). We will add results for multi-step forecasts, as well as the reverse overlapping case, in a future draft. In all cases, the primary DGPs are parameterized to roughly reflect the properties (estimated with 1961-2003 data from the 2007:Q1 vintage) of the change in the quarterly U.S. inflation rate (as measured by the GDP price index) and an output gap computed with the HP filter. The variable being forecast roughly corresponds to the change in inflation; the variables used to forecast inflation have properties similar to those of the HP output gap.

4.1 Monte Carlo design: non-nested case

In the non–nested forecast case, we consider a DGP that relates y to lags of y, x_1, and x_2, with x_1 independent of x_2 and both variables following AR processes. This basic VAR structure generates the final or true values of y, x_1, and x_2. One model used to forecast y includes lags of y and just x_1; the other includes lags of y and just x_2. To establish a baseline, we first consider the properties of forecast tests based entirely on the final data, with final data used to generate and evaluate forecasts (using the data y_t, $x_{1,t}$, and $x_{2,t}$ generated as detailed below). Data are generated using independent draws of innovations from the standard normal distribution and the autoregressive structure of the DGP. The complete data generating process for the final data takes the following form.

$$\begin{aligned}
y_t &= -.4y_{t-1} - .3y_{t-2} + .25x_{1,t-1} + (.25+\beta)x_{2,t-1} + u_{y,t} \qquad (3)\\
x_{1,t} &= 1.1x_{1,t-1} - .3x_{1,t-2} + u_{x1,t}\\
x_{2,t} &= 1.1x_{2,t-1} - .3x_{2,t-2} + u_{x2,t}\\
Var\begin{pmatrix} u_{y,t} \\ u_{x1,t} \\ u_{x2,t} \end{pmatrix} &= \begin{pmatrix} 1.0 & & \\ 0 & .5 & \\ 0 & 0 & .5 \end{pmatrix}
\end{aligned}$$

The coefficient β is set to zero in size experiments and 0.5 in power experiments.

We focus, of course, on data subject to revision, supposing the final values are released with a delay. In practice, data such as GDP are subject to many revisions. In the case of GDP-related data, three estimates are published 1, 2, and 3 months after the end of a quarter; subsequent estimates are published in three annual revisions; and yet further revisions are published in periodic benchmark revisions. In our Monte Carlo exercises, we try to simplify matters while at the same time preserving some of the essential features of actual revisions. In size experiments, we allow for two revisions of an initially published estimate. 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 $x_{2,t}(t)$). Updated estimates ($y_t(t+1)$, $x_{1,t}(t+1)$, and $x_{2,t}(t+1)$) are published in period $t+1$. The final estimates (y_t, $x_{1,t}$, and $x_{2,t}$) are treated as being published in period $t+8$. The particular dating is of course arbitrary, but our intention is to capture the empirical regularity of a combination of early revisions and late revisions. The first revision (published in $t+1$ in our simplified dating) is meant to correspond to the first revision of NIPA estimates; the second is meant to correspond to revisions of NIPA data published two years after the initial estimate (published in $t+8$ in our simplified dating).

In power experiments, we simplify things a bit in the interest of better controlling which forecast is more accurate in real time, and consider a single revision, with initial estimates published in period t and final estimates published in period $t+4$. Motivated by work in such studies as Croushore and Stark (2003), Faust and Wright (2005), and Arouba (2006) on predictability in data revisions, the revision processes have a common general structure, relating a revision between the prior estimate and current estimate to the prior estimate and an independent innovation.

The data generating process for the revisions incorporated in the size experiments is given by the following.

$$\begin{aligned}
y_t - y_t(t+1) &= \gamma_{y,1} y_t(t+1) + v_{1,y,t} \\
x_{1,t} - x_{1,t}(t+1) &= \gamma_{x1,1} x_{1,t}(t+1) + v_{1,x1,t} \\
x_{2,t} - x_{2,t}(t+1) &= \gamma_{x2,1} x_{2,t}(t+1) + v_{1,x2,t} \\
y_t(t+1) - y_t(t) &= \gamma_{y,2} y_t(t) + v_{2,y,t} \\
x_{1,t}(t+1) - x_{1,t}(t) &= \gamma_{x1,2} x_{1,t}(t) + v_{2,x1,t} \\
x_{2,t}(t+1) - x_{2,t}(t) &= \gamma_{x2,2} x_{2,t}(t) + v_{2,x2,t}.
\end{aligned} \tag{4}$$

In implementation, we generate the preliminary data with a simple iterative approach: from the final data and draws of the errors $v_{1,y,t}$, $v_{1,x1,t}$, and $v_{1,x2,t}$, we use the first three equations in (4) to construct the second estimates, published in period $t+1$; from the second estimates and draws of the errors $v_{2,y,t}$, $v_{2,x1,t}$, and $v_{2,x2,t}$, we use the last three equations (4) to construct the initial estimates, published in period t. In the power experiments, we allow for a single revision, with the final estimate published with a four-period delay:

$$\begin{aligned}
y_t - y_t(t) &= \gamma_{y,1} y_t(t) + v_{1,y,t} \\
x_{1,t} - x_{1,t}(t) &= \gamma_{x1,1} x_{1,t}(t) + v_{1,x1,t} \\
x_{2,t} - x_{2,t}(t) &= \gamma_{x2,1} x_{2,t}(t) + v_{1,x2,t}
\end{aligned} \tag{5}$$

Our parameterizations of the revision processes are drawn from empirical estimates for real-time U.S. data on the change in GDP inflation and the HP output gap from 1965 through 2003. As detailed in section 5, this empirical evidence is for revisions from first estimates to second (from the first vintage in the Philadelphia Fed's real time data set to the second) and for revisions from the second estimate to one published two years after the initial estimate. The innovations in the revisions equations are all iid normal random variables. In size experiments, the standard deviations of the innovations $v_{1,y,t}$, $v_{1,x1,t}$, $v_{1,x2,t}$, $v_{2,y,t}$, $v_{2,x1,t}$, and $v_{2,x2,t}$ are, respectively, 0.9, 1.3, 1.3, 0.5, 0.2, and 0.2. In power experiments, the standard deviations of $v_{1,y,t}$, $v_{1,x1,t}$, and $v_{1,x2,t}$ are, respectively, 0.5, 0.2, and 0.2. In the baseline size experiments with predictable revisions, the γ coefficients are $\gamma_{y,1}$ = -0.2, $\gamma_{x1,1}$ = -0.3, $\gamma_{x2,1}$ = -0.3, $\gamma_{y,2}$ = -0.05, $\gamma_{x1,2}$ = -0.2, and $\gamma_{x2,2}$ = -0.2. In the baseline power experiments with predictable revisions, the γ coefficients are $\gamma_{y,1}$ = -0.2, $\gamma_{x1,1}$ = -0.3, and $\gamma_{x2,1}$ = -0.3.

In all non–nested experiments, we test for equal accuracy of two forecasts, from the following models:

$$y_{t+1} = a_0 + a_1 y_t + a_2 y_{t-1} + a_3 x_{1,t} + u_{1,t+1} \quad (6)$$

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 x_{2,t} + u_{2,t+1}. \quad (7)$$

At each forecast origin $t = R, ..., T$, the parameters of the forecasting models are estimated recursively by OLS. In the size experiments, at each forecast origin t, the observable time series for each variable consists of an initial or first vintage estimate for period t, second vintage estimates for periods $t-1$ through $t-7$, and final values for periods $t-8$ and earlier. In power experiments, the observable time series as of period t consist of initial vintage estimates for periods t through $t-3$ and final values for periods $t-4$ and earlier. However, in experiments without data revisions, the data samples consist solely of the final data. As forecasting moves forward in time, the models are re-estimated with an expanding sample of data.

In evaluating forecasts, we compute forecast errors using actual values of y taken to be the initial estimate published in period t, $y_t(t)$ (except that we use the final value y_t in experiments without data revisions). In the size experiments, using the second estimate published in period $t+1$ yields very similar results. We form two versions of the MSE-t test, one with a standard error of just an estimate of S_{dd} and the other with an estimate of $S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$. In this draft, with the forecasts limited to the one–step horizon, all HAC estimates in these variances are computed imposing a bandwidth of 0. Both 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 (using a critical value of ±1.96).

Finally, with quarterly data in mind, we consider a range of sample sizes. For simplicity, we report results for a single R setting, of 80; other practical settings of R yield similar results. We report results for four different P settings: 20, 40, 80, and 160 (corresponding to π values of 0.2, 0.5, 1, and 2).

4.2 Monte Carlo design: nested case

In light of the possibility of a singularity in the variance Ω in the nested case, we begin our nested model forecast analysis with a DGP simple enough that it allows us to easily determine whether, even in the presence of predictable revisions, Ω is zero. In this simple case, the process generating the final data is as follows:

$$\begin{aligned} y_t &= u_{y,t} && (8) \\ x_t &= 3 + u_{x,t} \\ u_{y,t}, u_{x,t} & \quad \text{iid } N(0,1). \end{aligned}$$

The variable x is never revised. The predictand y is subject to one revision: $y_t = \kappa + y_t(t) + v_t$, published one period after the initial estimate. To make the revisions predictable, we use $\kappa = 1$. With this simple DGP, we first run simulations designed to verify our basic theoretical results, using a setup that we can analytically verify as falling within our assumptions. In these simulations, the null model of y is $y_t = u_{1,t}$; the forecast is then simply 0 (which reduces $-JB_2J' + B_2$ to just B_2). The alternative forecasting model is $y_t = a + bx_t + u_{2,t}$. We then consider an alternative setup in which we can analytically verify that $F \neq 0$, but that $\Omega = 0$. In these simulations, the null model of y is $y_t = a + u_{1,t}$; the alternative is $y_t = a + bx_t + u_{2,t}$.

We also examine DGPs with a structure similar to that used in the non–nested case. These DGPs relate y to lags of y and x, with x following an AR process. This basic VAR structure generates the final or true values of y and x. The null model used to forecast y includes just lags of y; the alternative model includes lags of y and x. To establish a baseline, we first consider the properties of forecast tests based entirely on the final data. Data are generated using independent draws of innovations from the standard normal distribution and the autoregressive structure of the DGP. More specifically, the complete data generating process for the final data takes the following form.

$$\begin{aligned} y_t &= -.4y_{t-1} - .3y_{t-2} + \beta_3 y_{t-3} + \beta_4 y_{t-4} + \beta_{22} x_{t-1} + u_{y,t} && (9) \\ x_t &= 1.1x_{t-1} - .3x_{t-2} + u_{x,t} \\ Var\begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix} &= \begin{pmatrix} 1.0 & \\ .25 & .5 \\ & \end{pmatrix} \end{aligned}$$

In size experiments, we use one DGP (NDGP 2) with β_3 = -0.2, β_4 = 0.1, and $\beta_{22} = 0$, and another (NDGP 3) with $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_{22} = 0$. We also consider versions of these DGPs in which the residual in the y equation follows a GARCH process, parameterized to keep the unconditional variance the same as in the conditionally homoskedastic parameterization:

$$u_{y,t} = \sqrt{s_t}\tilde{u}_{y,t}, \quad \tilde{u}_{y,t} \text{ iid } N(0,1) \tag{10}$$

$$s_t \quad = \quad .1 + .6 s_{t-1} + .3 u_{y,t-1}^2$$

In power experiments, we report results for just one conditionally homoskedastic DGP, using β_3 = -0.2, $\beta_4 = 0.1$, and $\beta_{22} = 0.3$ (as noted above, these parameterizations are drawn from empirical estimates with inflation and an output gap). The forecasting models vary across these experiments, as described below.

We focus on data subject to revision, supposing the final values are released with a delay. We model and generate the revisions as we did for the non–nested case.

$$\begin{aligned} y_t - y_t(t+1) &= \gamma_{y,1} y_t(t+1) + v_{1,y,t} \qquad (11) \\ x_t - x_t(t+1) &= \gamma_{x,1} x_t(t+1) + v_{1,x,t} \\ y_t(t+1) - y_t(t) &= \gamma_{y,2} y_t(t) + v_{2,y,t} \\ x_t(t+1) - x_t(t) &= \gamma_{x,2} x_t(t) + v_{2,x,t} \end{aligned}$$

In implementation, we generate the preliminary data with a simple iterative approach: from the final data and draws of the errors $v_{1,y,t}$ and $v_{1,x,t}$, we use the first two equations in (11) to construct the second estimates, published in period $t+1$; from those second estimates and draws of the errors $v_{2,y,t}$ and $v_{2,2,t}$, we use the last two equations (11) to construct the initial estimates, published in period t. In the power experiments, we allow for a single revision, with the final estimate published with a four-period delay:

$$\begin{aligned} y_t - y_t(t) &= \gamma_{y,1} y_t(t) + v_{1,y,t} \qquad (12) \\ x_t - x_t(t) &= \gamma_{x,1} x_t(t) + v_{1,x,t} \end{aligned}$$

Our parameterizations of the revision processes are the same as in the non–nested case, drawn from empirical estimates for real-time U.S. data on the change in GDP inflation and the HP output gap from 1965 through 2003. The innovations in the revisions equations are all iid normal random variables. In size experiments, the standard deviations of the innovations $v_{1,y,t}$, $v_{1,x,t}$, $v_{2,y,t}$, and $v_{2,x,t}$ are, respectively, 0.9, 1.3, 0.5, and 0.2. In power experiments, the standard deviations of $v_{1,y,t}$ and $v_{1,x,t}$ are, respectively, 0.5 and 0.2. In the baseline size experiments with predictable revisions, the γ coefficients are $\gamma_{y,1}$ = -0.2, $\gamma_{x,1}$ = -0.3, $\gamma_{y,2}$ = -0.05, and $\gamma_{x,2}$ = -0.2. In the baseline power experiments with predictable revisions, the γ coefficients are $\gamma_{y,1}$ = -0.2 and $\gamma_{x,1}$ = -0.3.

In experiments for the DGP we refer to as NDGP 2, we consider forecasts from models of the form

$$y_{t+1} = a_0 + a_1 y_t + a_2 y_{t-1} + a_3 y_{t-2} + a_4 y_{t-3} + u_{1,t+1} \tag{13}$$

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 y_{t-2} + b_4 y_{t-3} + b_5 x_t + u_{2,t+1}. \tag{14}$$

In experiments for NDGP 3, the forecasting models are

$$y_{t+1} = a_0 + a_1 y_t + a_2 y_{t-1} + u_{1,t+1} \tag{15}$$

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 y_{t-2} + b_4 y_{t-3} + b_5 x_t + u_{2,t+1}. \tag{16}$$

At each forecast origin $t = R, ..., T$, the parameters of the forecasting models are estimated recursively by OLS. In the size experiments with NDGP 1, at each forecast origin t, the observable time series for y consists of an initial or first vintage estimate for period t and final estimates for all prior periods; all of the observable data for x are the final data. In the size experiments with NDGP 2 and NDGP 3, at each forecast origin t, the observable time series for each variable consists of an initial or first vintage estimate for period t, second vintage estimates for periods $t-1$ through $t-7$, and final values for periods $t-8$ and earlier. In power experiments, the observable time series as of period t consist of initial vintage estimates for periods t through $t-3$ and final values for periods $t-4$ and earlier. However, in experiments without data revisions, the data samples consist solely of the final data. As forecasting moves forward in time, the models are re-estimated with an expanding sample of data.

In evaluating forecasts, we compute forecast errors using actual values of y_t taken to be the initial estimate published in period t, $y_t(t)$ (except that we use the final value y_t in experiments without data revisions). In the size experiments with NDGP 2 and NDGP 3, using the second estimate published in period $t+1$ yields very similar results. The null hypothesis is that the variables included in the larger model and not the smaller have no predictive content. To test this null, from the forecast errors we form the MSE-F test and various versions of the MSE-t test, and compare them to various sources of critical values. We reject the null if the test statistic exceeds the relevant right-tail critical value (i.e., in the nested case, we conduct one-sided tests). We report the percentage of 10,000 simulations in which the null of equal accuracy is rejected at the 5% significance level.

More specifically, under $\pi > 0$ asymptotics, we construct the test statistic MSE-$F = P(MSE_1 - MSE_2)/MSE_2$ and compare it against asymptotic critical values from McCracken (2006). We construct the conventional version of the MSE-t test, defined as

MSE-t $= P^{1/2}(MSE_1 - MSE_2)/\hat{S}_{dd}^{1/2}$ (note that $\hat{S}_{dd} = \hat{\Gamma}_{dd}(0)$ for $\tau = 1$), and compare it against critical values from McCracken (2006). We construct two other versions of the MSE-t test based on alternative standard errors and compare them against standard normal critical values. These two versions use the following standard errors: the square root of $\widehat{\Omega} = 2\hat{\Pi}\widehat{F}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{F}'$ (note that the formula simplifies in the NDGP 1 experiments because model 1 has no estimated parameters) and the square root of $\widehat{\Omega} = \hat{S}_{dd} + 2\hat{\Pi}\widehat{F}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{F}'$.

Under $\pi = 0$ asymptotics, we compare the MSE-F test multiplied by $(R/P)^{1/2}$ against ($\pi = 0$) critical values from McCracken. We also compare the conventional MSE-t test computed with the square root of $\hat{S}_{dd}$ and an MSE-t test using a standard error given by the square root of $\widehat{\Omega} = \widehat{F}(-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{S}_{hh}\ (-J\widehat{B}_1 J' + \widehat{B}_2)\widehat{F}'$ against standard normal critical values.

In all cases we estimate the pieces of the variance of the MSE differential as discussed in the previous section, imposing the absence of serial correlation implied by the one–step horizon, such that the long-run covariances $\hat{S}_{dd}$, $\widehat{S}_{hh}$, and $\widehat{S}_{dh}$ are estimated with just $\hat{\Gamma}_{dd}(0)$, $\hat{\Gamma}_{hh}(0)$, and $\hat{\Gamma}_{dh}(0)$, respectively.

Finally, with quarterly data in mind, we consider a range of sample sizes. In NDGP 1 experiments, to best check our theory results, we use combinations of P and R that yield π values from quite small to quite large (smaller and larger than is typical in macro forecasting studies): $R = 100$ with $P = 10$, 40, 100, and 400. In NDGP 2 and NDGP 3 experiments, we use $R = 80$ with $P = 20$, 40, 80, and 160 (corresponding to π values of 0.2, 0.5, 1, and 2). Again, our intention is to focus on sample sizes most relevant for real time macroeconomic forecast analysis. Other practical settings of R yield similar results.

4.3 Monte Carlo results: non-nested case

Table 1 reports size results from simulations of various versions of NNDGP 1, varying in the extent to which the data are revised.[5] As a benchmark, we begin with the case in which the data are not revised at all (first panel). Consistent with other evidence in the literature, in this case, a conventional t-test for equal MSE (MSE-t using S_{dd}) tends to be slightly oversized, with size ranging from 6 to 7 percent. With no revisions, the adjustment term $2\Pi(FBS_{dh} + FBS_{hh}BF')$ limits to zero, but in small samples tends to be small and

[5] The random numbers are generated such that the final data in the no revisions experiment are the same as the final data in the experiment in which all variables are subject to noisy revisions.

positive, causing the adjusted t-test to be modestly to slightly undersized, with size between 3 and 4 percent.

Consider now the size of the tests in the case of predictable revisions in all variables (second panel). 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 (understate) the variance impact of the predictable revisions. However, in practice, the unadjusted test's size is 5 percent for most sample sizes and 7 percent for $P = 160$. Incorporating the adjustment called for by the asymptotic results in section 3 causes the test to be significantly undersized, with size at 1 percent for most samples and 2 percent for $P = 160$. Some supplemental simulations with larger sample sizes indicate these outcomes reflect small sample properties, not a problem with the asymptotics. In simulations with $R = 320$ and $P = 60$, 120, 240, and 480, the size of the adjusted test is 4 to 5 percent; the size of the unadjusted test ranges from 7 to 11 percent. In simulations with $R = 800$ and $P = 200$, 400, 800, and 1600, the adjusted test is correctly sized, while the size of the unadjusted test ranges from 9 to 20 percent.

Table 2 provides power results from simulations of NNDGP 1. In the benchmark case of no data revisions (first panel), the unadjusted and adjusted MSE-t tests have very similar power, ranging from 26 (unadjusted) vs. 29 (adjusted) percent for $P = 20$ to 99 percent (both) for $P = 160$. Introducing noisy revisions to both the y and x variables significantly lowers power, but very similarly for the unadjusted and adjusted test statistics, such that it remains the case that the powers of the two tests are quite similar. For example, with $P =$ 80, the unadjusted and adjusted MSE-t tests have power of 50 and 53 percent, respectively.

4.4 Monte Carlo results: nested case

In the nested forecast model case, we begin by using the very simple NDGP 1 to assess the practical relevance of our theoretical results. Table 3 provides results for two versions of the DGP, the first of which (left panel of numbers) uses a null model that relates y to just an error, so that the null forecast is 0. Consistent with our theoretical results, with noisy revisions in NDGP 1, the standard MSE-t and MSE-F statistics compared against critical values from McCracken (2006) suffer huge size distortions, with size approaching the 50% level predicted by the theory. Comparing a conventional MSE-t test using S_{dd} as the variance estimate against standard normal critical values also yields significant oversizing, with size ranging from 21 ($P = 10$) to 43 ($P = 400$) percent. However, comparing the MSE-t

test using $2\Pi F(-JB_1J'+B_2)S_{hh}(-JB_1J'+B_2)F'$ as the variance (assuming $\pi > 0$) against standard normal critical values yields roughly correct inferences, with nominal size at 6 percent. Similarly, the MSE-t test that uses $S_{dd}+2\Pi F(-JB_1J'+B_2)S_{hh}(-JB_1J'+B_2)F'$ as the variance estimate is about correctly sized for all but the smallest sample size. This result, too, confirms the asymptotic result that, with the predictable revisions, the conventional variance S_{dd} has a population value of 0. Finally, as might be expected, using the variance implied by the $\pi = 0$ asymptotics ($FBS_{hh}BF'$) works reasonably well with $P = 10$, yielding size of 6 percent, but more significant undersizing as P rises (relative to R).

To illustrate what happens when a singularity in Ω causes our asymptotic results to break down, the right panel of Table 3 reports results from an experiment in which the null forecasting model relates y to a constant and an error (so that the null forecast is the mean of y), rather than to just an error (so the null forecast is 0) as in the previous results. As noted above, in this case, it can be shown analytically that, although $F \neq 0$ (with noisy revisions), a singularity makes $\Omega = 0$. As a result, our proposed test may not be reliable, but there is no reason to expect conventional tests based on other asymptotics to be reliable, either. In the simulations, the most accurate test seems to be the MSE-F test compared against $\pi > 0$ critical values, with size ranging from 3 to 6 percent (however, using $\pi = 0$ asymptotics yields significant undersizing). The MSE-t test with conventional variance S_{dd} compared against McCracken's (2006) critical values is significantly oversized, with size between 11 and 15 percent. The same test compared against standard normal critical values is somewhat undersized (with size of 1 to 3 percent), except for very small P. Even though $\Omega = 2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ has a limiting value of 0, in finite samples the MSE-t test using this variance estimate is a bit less oversized than the conventional MSE-t test compared against McCracken's critical values. The size of our proposed test ranges from 9 to 11 percent. Accordingly, in finite samples, our asymptotic approximation for the MSE-t test doesn't seem to be materially worse than any other in the event a singularity in the relevant variance renders all existing asymptotics invalid.

Moving on to more empirically relevant DGPs, Table 4 provides size results from NDGP 2 and NDGP 3. Again, as a benchmark, we consider experiments with no data revisions (top panel). Consistent with results from our prior work, in this setting the MSE-F test and MSE-t test based on S_{dd} compared against $\pi > 0$ critical values from McCracken (2006) range from correctly sized to slightly oversized. For example, with NDGP 3, the sizes of these two tests vary from 4 to 6 percent. However, comparing the MSE-F test against $\pi = 0$ critical values yields significant undersizing. Moreover, consistent with our prior work and results in other studies such as Clark and West (2007), the conventional MSE-t test based on S_{dd} and standard normal critical values is also significantly undersized, with size ranging from 0 to 3 percent. Our proposed MSE-t test using a variance of $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ and standard normal critical values, which has no asymptotic justification in the absence of data revisions, ranges from significantly undersized (NDGP 3, $P = 160$) to oversized (NDGP 2, $P = 20$). Not surprisingly, given the preceding results, the MSE-t test using a variance of variance of $S_{dd}+2\Pi F(-JB_1J'+B_2)S_{hh}(-JB_1J'+B_2)F'$ and standard normal critical values yields a consistently and significantly undersized test, essentially never rejecting the null.

With noisy revisions in NDGP 2 and NDGP 3 (second panel of Table 4), the standard MSE-t and MSE-F statistics compared against critical values from McCracken (2006) range from being about correctly sized (MSE-t in NDGP 3, $P = 20$, 40, and 60) to significantly oversized (MSE-F in NDGP 2, all P). The conventional MSE-t test with S_{dd} as the variance estimate and standard normal critical values is consistently undersized, with size ranging from 1 to 3 percent. The performance of our proposed MSE-t test using $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ as the variance (assuming $\pi > 0$) and standard normal critical values is mixed, ranging from undersized (NDGP 3, $P = 160$) to oversized (NDGP 2, all P). Using a variance estimate that sums S_{dd} with $2\Pi F(-JB_1J'+B_2)S_{hh}(-JB_1J'+B_2)F'$ again yields a poor result, essentially never rejecting the null. Admittedly, these results suggest some practical limitations to the asymptotic results in section 3. Our asymptotic theory implies the standard MSE-t and MSE-F statistics compared against critical values from McCracken (2006) and MSE-t test with S_{dd} compared against standard normal critical values should be oversized, unless a singularity makes Ω equal to 0. In contrast, the MSE-t test with variance $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ should be correctly sized. In broad terms, the results in the second panel are consistent with these implications, but clearly the performance of the conventional tests is better and that of our proposed test worse than the theory implies. One explanation could be that the asymptotic approximations are imperfect guides in finite samples. Another could be that the F matrix is not non-zero, but it certainly is with predictable revisions in NDGP 2 and NDGP 3. Another explanation could be that, for these DGPs, Ω is nonetheless 0 (large sample simulations seem to rule this out for NDGP 3, but suggest it could be the case for NDGP 2).

Because conditional heteroskedasticity should reduce the potential for singularities in Ω (with predictable revisions), the last panel of Table 5 reports results for DGPs with GARCH in the innovations of the y equation. However, the introduction of GARCH produces little change in results relative to the baseline with conditional homoskedasticity in all innovations. In all cases, sizes are very similar across panels 2 and 3, with rejection rates in the GARCH panel at most a percentage point different from those in the conditional homoskedasticity (second) panel.

Table 5 provides power results from NDGP 3. Results for the benchmark case of no data revisions (top panel) are in line with previous results in the literature. The MSE-F test compared against asymptotic critical values (for $\pi > 0$) from McCracken (2006) is generally more powerful than the MSE-t test using S_{dd} and McCracken's critical values, which is in turn considerably more power than the same MSE-t test compared against standard normal critical values. Despite having no asymptotic justification in the absence of data revisions, our proposed MSE-t test with variance $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ and standard normal critical values is at least as powerful as the other MSE-t variants, and comparable in power to the MSE-F test. For example, with $P = 40$, the powers of the MSE-F, conventional MSE-t against McCracken critical values, and adjusted MSE-t test against standard normal critical values are, respectively, 71, 51, and 65 percent. As would be expected in light of the relatively poor power of MSE-t with S_{dd} and standard normal critical values, the variant using a variance of $S_{dd}+2\Pi F(-JB_1J'+B_2)S_{hh}(-JB_1J'+B_2)F'$ has especially poor power, at 20 percent in the same example.

With predictable revisions (second panel of Table 5), the powers of all tests are significantly lower than in the benchmark case of no revisions. It remains the case that the MSE-F test compared against McCracken's (2006) critical values (for which power ranges from 20 to 70 percent) is more powerful than a MSE-t test using S_{dd} and McCracken's critical values (for which power ranges from 12 to 63 percent). However, the relative power of the adjusted MSE-t test with variance $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ and standard normal critical values is better than in the no-revision baseline. For the $P = 20$ and 40 samples, this test is significantly more powerful than MSE-F (with $\pi > 0$), rejecting the null 15-17 percent more frequently than the MSE-F test does. As P rises, though, the advantage of the adjusted MSE-t dissipates, and even reverses for $P = 160$. Finally, data revisions make the power of the conventional MSE-t test with variance S_{dd} compared against standard normal critical values abysmal, ranging from just 9 to 21 percent.

4.5 Monte Carlo summary

Overall, the Monte Carlo analysis confirms that, in practical applications, conventional testing approaches that ignore the impact of predictable revisions can lead to incorrect inferences. Our proposed testing approach that takes revisions into account can lead to more reliable inferences, although not certainly so. In the non–nested case, an unadjusted t-test for equal MSE modestly over-rejects the null. Our proposed adjusted test tends to under-reject the null. Therefore, in practical applications with non–nested forecasts, it is probably useful to consider both tests. In the nested case, the conventional MSE-F and MSE-t tests compared against critical values from McCracken (2006) range from nearly correctly sized to modestly to significantly oversized. The conventional MSE-t test compared against standard normal critical values is severely undersized and has very low power. The performance of our proposed adjusted MSE-t test seems to be mixed, as the test ranges from modestly undersized to modestly oversized. Overall, for practical applications with nested forecasts, it is probably a good idea to consider our proposed test in conjunction with the conventional MSE-F and MSE-t tests compared against critical values from McCracken (2006).

5 Application to Inflation Forecasting

In this section we use the tests and inference approaches described above to determine whether, in real time data, various measures of economic activity have predictive content for inflation. The inflation measure we forecast is the change in the inflation rate of the GDP price index. We consider one–quarter and one–year ahead forecasts of inflation from an AR model, a model including lags of the change in inflation and GDP growth, and a model including lags of the change in inflation and the output gap (HP detrended output). To illustrate non–nested testing, we compare forecasts from the model with GDP growth to the model with the output gap. To illustrate nested testing, we compare forecasts from the model with GDP growth to the AR model. Real-time evidence in Orphanides and van Norden (2005) suggests that GDP growth may be superior to the output gap for forecasting inflation, as well as to the AR model.

5.1 Data

Data on real output and the price index are taken from the Federal Reserve Bank of Philadelphia's Real–Time Data Set for Macroeconomists (RTDSM). For simplicity, we simply use the notation "GDP" and "GDP price index" to refer to the output and price series, even though the measures are based on GNP and a fixed weight deflator for much of the sample. The full forecast evaluation period runs from 1970:Q1 through 2003:Q4. As described in Croushore and Stark (2001), the vintages of the RTDSM are dated to reflect the information available around the middle of each quarter. Normally, in a given vintage t, the available NIPA data run through period $t-1$.[6] For each forecast origin t in 1970:Q1 through 2003:Q4, we use the real time 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 1961:1+$\tau - 1$, where τ denotes the forecast horizon.

In evaluating real time forecast accuracy, we consider a range of possible definitions (vintages) of actual inflation. One estimate is the first one available in the RTDSM, one quarter after the end of the forecast observation date (i.e., inflation for period t published in period $t+1$). Another is the second estimate or vintage available in the RTDSM, published with a two–quarter delay. Studies such as Romer and Romer (2000) use the second available estimates of the GDP/GNP deflator as actuals in evaluating forecast accuracy. We also consider estimates of inflation published with delays of five and 13 periods.

5.2 Models

Following Stock and Watson (1999, 2003), among many others, we treat inflation as being close enough to I(1) to warrant imposing a unit root and compare forecasts of the change in inflation from Phillips curve specifications including different measures of economic activity (GDP growth or the output gap) to forecasts from a simple autoregressive model. We report forecast results for the two horizons that seem to be most widely used in previous studies and most interesting to policymakers: one quarter and one year.

Letting τ denote the forecast horizon (in quarters), we use reduced–form Phillips curves

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{L-1} \alpha_l \Delta\pi_{t-l} + \beta x_t + u_{PC,t+\tau}, \tag{17}$$

[6] In the case of the 1996:Q1 vintage, with which the BEA published a benchmark revision, the data run through 1995:Q3 instead of 1995:Q4.

where inflation is $\pi_t^{(\tau)} \equiv (400/\tau)\ln(p_t/p_{t-\tau})$, $\pi_t^{(1)} \equiv \pi_t$, and x_t is a measure of economic activity. The same basic model specification has been used in studies such as Stock and Watson (1999, 2003) and Clark and McCracken (2006). In one version of this model, the x_t variable is defined as the four–quarter GDP growth rate, $100 \ln(\text{GDP}_t/\text{GDP}_{t-4})$. In the other, x_t is defined as (100 times) HP-detrended log GDP. In both models, the lag order L is 4.

In addition to comparing forecasts from one version of (17) with GDP growth to another with the output gap, we compare forecasts from the model with GDP growth to forecasts from an AR model for the change in inflation. Following the aforementioned studies, this AR model takes the form

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{M-1} \alpha_l \Delta\pi_{t-l} + u_{AR,t+\tau}. \tag{18}$$

We use an AR model lag order of $M = 2$.

In computing the various versions of the MSE-t test, we use the Newey and West (1987) estimator of the long-run variances S_{dd}, S_{dh}, and S_{hh}, with a bandwidth of $2(\tau-1)$ (following Cochrane (1991)).

5.3 Results

As a first step, in Table 6 we document the predictability of revisions to (quarterly) GDP growth, the HP output gap, and changes in GDP inflation. Following Croushore and Stark (2003), we report correlations of various revisions to the variable of interest with various vintages of estimates. For example, the first element of each block provides the correlation of (1) the second available estimate of the variable in question less the first estimate of the variable (the first revision) with (2) the first available estimate of the variable. A negative correlation of a revision with an estimate available at the time of the baseline estimate in the revision means the revision is predictable. The reported correlations are based on a sample period of 1965:Q4 through 2003:Q4. Our correlations for GDP growth, shown in the top block, are quite similar to those in Croushore and Stark (2003), suggesting some noise component in real time GDP growth estimates. Our estimates for the HP output gap, given in the middle block, point to somewhat stronger predictability of revisions to the HP output gap. For example, the correlation of the first revision of the gap with the first estimate of the gap is -.87; the correlation of the revision from the second estimate to the estimate available two years later with the second estimate of the output gap is -.30. Estimates for the change in GDP inflation in the last block also point to some predictability of revisions. The correlation of the first revision of the inflation change with the first estimate of the change are -.15; the correlation of the revision from the second estimate to the estimate available two years later with the second estimate of (the change in) inflation is -.45.

Table 7 presents results for the (non–nested) comparison of forecasts from the models with GDP growth (model 1) and the output gap (model 2). For most samples and definitions of actuals, although not all, the model with GDP growth yields slightly more accurate forecasts. The advantage of the model with GDP growth is considerably greater in year-ahead forecasts than one quarter-ahead forecasts. However, there is little evidence of statistical significance in the forecast accuracy differences. If the conventional variance S_{dd} is used in forming the t-test, the null of equal accuracy is rejected only once at the one–step horizon (for the 1985-2003 sample using the inflation estimates published with a 13 period delay as actuals), but for all 1970-2003 and 1985-2003 samples at the one–year ahead horizon. Consistent with our theory and Monte Carlo evidence, in most cases, taking account of the potential for predictability in the data revisions raises the estimated standard error. At the one–step horizon, though, the impact is pretty small in most cases, particularly in results for the 1970-2003 and 1970-84 samples. At the one–step horizon, the adjustment has a bigger impact in the 1985-2003 results. Most notably, for the 1985-2003 sample using the inflation estimates published with a 13 period delay as actuals, the null of equal accuracy is not rejected based on the adjusted variance estimate, but it is when based on the unadjusted variance. The adjustment has a considerably bigger impact in the one–year ahead forecasts. The rejections of equal accuracy for all 1970-2003 and 1985-2003 samples based on the t-test using S_{dd} go away when the test uses the adjusted variance Ω.

Table 8 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, slightly so at the one–step horizon and more substantially at the one–year horizon. When we abstract from the potential impact of predictable data revisions on test behavior, and compare the MSE-F test and MSE-t test using S_{dd} to ($\pi > 0$) asymptotic critical values simulated as described in Clark and McCracken (2005), for most definitions of actual inflation we reject the null AR model with the full 1970-2003 sample of forecasts and the 1985-2003 sample.[7] At the one-year horizon, the null is also always

[7] Because of the serial correlation and possible heteroskedasticity in the forecast errors, we use the Monte

rejected for the 1970-84 sample. If the same MSE-t test is compared against standard normal critical values (1.282 for a one-sided 10% test), for one–step ahead forecasts the null is consistently rejected for the 1985-2003 sample but never rejected for the 1970-2003 period. For one–year ahead forecasts, the null AR model is nearly always rejected for the 1970-2003 and 1970-84 samples, but never for the 1985-2003 period. However, in the presence of data revisions, the tests based on McCracken's asymptotics are prone to over-rejecting the null; the conventional MSE-t compared against the standard normal generally under-rejects. Taking account of data revisions by using a variance of $\Omega = 2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$ in the MSE-t test increases the (absolute) value of the t–statistic in all but one case. However, at the one–step horizon, in only two cases — forecasts for 1970-2003 and 1970-84 evaluated with first available estimates of inflation — is the adjusted t–statistic significant when the unadjusted t–statistic (compared against standard normal critical values) is not. At the one–year horizon, using the adjusted standard error has a big impact on inference for the 1985-2003 sample, with the adjusted t-test rejecting the null and the unadjusted test not rejecting for three of the four definitions of actual inflation.

Overall, at the one–year ahead horizon, the adjusted MSE-t test confirms the strong evidence in favor of the model with GDP growth for all samples. At the one-step horizon, adjusted MSE-t test confirms the strong evidence in favor of the model with GDP growth for the 1985-2003 sample. However, the adjusted test indicates the evidence in favor of the model for the 1970-2003 sample to be much weaker than do the tests based on McCracken's (2006) asymptotics. For the 1970-2003 sample, the adjusted t-test rejects the null only once, with the first available definition of actuals; the tests compared against critical values from McCracken's asymptotics reject the null for all four definitions of actuals.

6 Conclusion

In this paper we derive the limiting distributions for tests of equal predictive ability when forecasting with real time vintage data. Specifically, we address the impact of revisions exhibiting news and noise on the asymptotic distributions of the t–statistic for equal MSE between non-nested models developed by Diebold and Mariano (1995) and West (1996) and the F– and t–type tests of equal MSE between nested models developed in Clark and McCracken (2005) and McCracken (2006). We show that when revised data is used

Carlo method outlined in Clark and McCracken (2005) to compute critical values based on the Clark and McCracken (2005) and McCracken (2006) asymptotics for the MSE-F and MSE-t tests.

to construct and evaluate forecasts these tests typically do not have the same asymptotic distributions as when the data is never revised. With these new distributions in hand, we show how to conduct asymptotically valid inference. In the cases we consider, the tests are asymptotically standard normal and hence inference can be conducted using the relevant tables.

Using our asymptotics, we then conduct a range of Monte Carlo simulations to examine the finite–sample properties of the tests. Overall, these results broadly confirm our asymptotic approximations. In terms of size, ignoring the data revisions can produce oversized tests. Taking revisions into account by using our proposed tests can yield more reliable inferences. Data vintage also has an impact on the power of the tests. Typically, power is lower in data subject to revision than in data that are unrevised. The revisions drive a wedge between the properties of the dependent variable defining the predictive model and that used for evaluation. Depending on the exact relationships across vintages, predictive content for one vintage need not imply the same for another.

In the final part of our analysis, we illustrate the usage of our tests with an application to competing forecasts of U.S. inflation.

References

Anatolyev, Stanislav (2007). Inference about predictive ability when there are many predictors. manuscript, New Economic School, Moscow.

Armah, N.A.C., Swanson, N.R. (2006), "Predictive Inference Under Model Misspecification with an Application to Assessing the Marginal Predictive Content of Money for Output," in *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, D. Rapach and M. Wohar, eds., Elsevier, forthcoming.

Aruoba, S. Borağan (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., McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105:85-110.

Clark, T. E., McCracken, M. W. (2005). Evaluating Direct Multi-Step Forecasts. *Econometric Reviews* 24:369-404.

Clark, T.E., McCracken, M.W. (2006), "The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample Evidence," *Journal of Money, Credit, and Banking* 38:1127-1148.

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

Cochrane, J.H. (1991). "Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations," *Journal of Finance* 46:207-234.

Corradi, V., Swanson, N. R. (2002). A consistent test for nonlinear out–of–sample predictive accuracy. *Journal of Econometrics* 110:353-81.

Corradi, V., Swanson, N. R., Olivetti, C. (2001). Predictive ability with cointegrated variables. *Journal of Econometrics* 105:315-358.

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

Croushore, Dean and Tom Stark (2001), "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics* 105, 111-30.

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

Diebold, F. X., Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business* *and Economic Statistics* 13:253-263.

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

Giacomini and White (2006)

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

Koenig, E.F., S. Dolmas and J. Piger (2003). The use and abuse of real-time data in economic forecasting. *The Review of Economics and Statistics* 85:618-628.

Mankiw, N.G. and M.D. Shapiro (1986). News or noise: an analysis of GNP revisions. *Survey of Current Business* 66:20-25.

Mankiw, N.G., D.E. Runkle and M.D. Shapiro (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. (2006). Asymptotics for out–of–sample tests of causality. *Journal of Econometrics*, forthcoming.

Meese, R., Rogoff, K. (1988). Was it real? The exchange rate–interest differential relation over the modern floating–rate period. *Journal of Finance* 43:933-948.

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

Orphanides, A., 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.

Stock, J. H., Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics* 44:293-335.

Stock, J. H., 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. (2001). Tests for forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics* 19:29-33.

West, K. D. and M.W. McCracken (1998). Regression–based tests of predictive ability. *International Economic Review* 39:817-840.

Wooldridge, J.M. and H. White (1998). Central limit theorems for dependent heterogeneous processes with trending moments, in H. White ed., *Topics in Econometric Theory: The selected works of Halbert White*, (Edward Elgar, Cheltenham).

Table 1. Non-Nested Model Size Results, NNDGP 1

($R = 80$, nominal size = 5%)

test	variance	$P =$ 20	40	80	160
		no revisions			
MSE-t	S_{dd}	.07	.06	.06	.06
adj. MSE-t	$S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$	.03	.03	.04	.04
		predictable revisions			
MSE-t	S_{dd}	.05	.05	.05	.07
adj. MSE-t	$S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$	.01	.01	.01	.02

Notes:

1. The DGPs are defined in equations (3) and (4). The forecasting models are given in equations (6) and (7).
2. R defines the size of the sample used to generate the first forecast. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000.
3. The second column gives the variance estimator used in the test statistic. All test statistics are compared against standard normal critical values of ±1.96.

Table 2. Non-Nested Model Power Results, NNDGP 1

($R = 80$, nominal size = 5%)

test	variance	$P =$ 20	40	80	160
		no revisions			
MSE-t	S_{dd}	.29	.52	.82	.99
adj. MSE-t	$S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$	.26	.51	.83	.99
		predictable revisions			
MSE-t	S_{dd}	.17	.28	.50	.80
adj. MSE-t	$S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$	.16	.28	.53	.84

Notes:
1. The DGPs are defined in equations (3) and (5). The forecasting models are given in equations (6) and (7).
2. See the notes to Table 1.

Table 3. Nested Model Size Results, NDGP 1

($R = 100$, nominal size = 5%)

			null forecast = 0				**null forecast = mean**			
test	*variance*	*c.v.*	$P = 10$	40	100	400	$P = 10$	40	100	400
$\pi > 0$:										
MSE-F		M	.08	.24	.37	.47	.03	.03	.04	.06
MSE-t	S_{dd}	M	.27	.40	.45	.49	.12	.11	.12	.15
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.06	.06	.06	.06	.09	.10	.11	.11
adj. MSE-t	$S_{dd} + 2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.02	.04	.05	.05	.00	.00	.00	.00
$\pi = 0$:										
MSE-F		M	.00	.03	.18	.37	.00	.00	.00	.01
MSE-t	S_{dd}	N	.21	.33	.38	.43	.06	.03	.02	.01
adj. MSE-t	$FBS_{hh}BF'$	N	.06	.04	.03	.00	.08	.09	.08	.06

Notes:

1. The DGP is defined in equation (8). In the left panel of results, the null forecasting model is $y_t = u_{1,t}$ (so the null forecast is 0). In the right panel, the null forecasting model is $y_t = a + u_{1,t}$. In both cases, the alternative forecasting model is $y_t = a + bx_t + u_{2,t}$.
2. R defines the size of the sample used to generate the first forecast. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000.
3. The second column gives the variance estimator used in the MSE-t statistics. The matrix $\tilde{B}$ is shorthand for $-JB_1J' + B_2$. The third column indicates what critical value is used. An 'M' means the critical value is taken from McCracken (2006); an 'N' means the critical value is taken from the standard normal distribution (1.645). All tests are one-sided, with the null rejected if the statistic exceeds the right-tail critical value.

Table 4. Nested Model Size Results, NDGPs 2 and 3

(R = 100, nominal size = 5%)

test	variance	c.v.	P = 20	40	80	160	P = 20	40	80	160
$\pi > 0$:			**NDGP 2, no revisions**				**NDGP 3, no revisions**			
MSE-F		M	.07	.06	.05	.05	.06	.05	.05	.04
MSE-t	S_{dd}	M	.08	.06	.06	.05	.06	.05	.05	.04
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.09	.08	.07	.05	.05	.04	.03	.01
adj. MSE-t	$S_{dd}+2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.00	.00	.00	.00	.00	.00	.00	.00
$\pi = 0$:										
MSE-F		M	.01	.01	.01	.02	.00	.00	.00	.00
MSE-t	S_{dd}	N	.03	.02	.01	.00	.02	.01	.00	.00
adj. MSE-t	$FBS_{hh}BF'$	N	.08	.07	.05	.04	.04	.03	.02	.01
$\pi > 0$:			**NDGP 2, noise**				**NDGP 3, noise**			
MSE-F		M	.12	.11	.11	.13	.08	.07	.07	.08
MSE-t	S_{dd}	M	.07	.07	.08	.10	.05	.04	.06	.08
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.09	.09	.09	.08	.05	.04	.03	.02
adj. MSE-t	$S_{dd}+2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.00	.00	.00	.00	.00	.00	.00	.00
$\pi = 0$:										
MSE-F		M	.02	.03	.04	.05	.01	.01	.01	.02
MSE-t	S_{dd}	N	.03	.02	.01	.01	.02	.01	.01	.01
adj. MSE-t	$FBS_{hh}BF'$	N	.09	.08	.07	.05	.04	.03	.02	.01
$\pi > 0$:			**NDGP 2, GARCH**				**NDGP 3, GARCH**			
MSE-F		M	.13	.12	.12	.14	.09	.07	.07	.08
MSE-t	S_{dd}	M	.08	.07	.09	.11	.05	.04	.06	.07
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.10	.10	.09	.09	.05	.04	.03	.02
adj. MSE-t	$S_{dd}+2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.01	.00	.01	.00	.00	.00	.00	.00
$\pi = 0$:										
MSE-F		M	.02	.03	.05	.06	.01	.01	.01	.02
MSE-t	S_{dd}	N	.03	.02	.02	.01	.02	.01	.01	.01
adj. MSE-t	$FBS_{hh}BF'$	N	.09	.09	.08	.06	.04	.03	.02	.01

Notes:

1. The DGPs are defined in equations (9) and (11). The NDGP 2 forecasting models are given in equations (13) and (14). The NGDP 3 forecasting models are given in equations (15) and (16). In the last panel, the residuals of the y equation are generated to follow the GARCH(1,1) process given in equation (10).
2. R defines the size of the sample used to generate the first forecast. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000.
3. The second column gives the variance estimator used in the MSE-t statistics. The matrix $\tilde{B}$ is shorthand for $-JB_1J'+B_2$. The third column indicates what critical value is used. An 'M' means the critical value is taken from McCracken (2006); an 'N' means the critical value is taken from the standard normal distribution (1.645). All tests are one-sided, with the null rejected if the statistic exceeds the right-tail critical value.

Table 5. Nested Model Power Results, NDGP 3

($R = 100$, nominal size = 5%)

test	variance	c.v.	$P = 20$	40	80	160
$\pi > 0$:			**NDGP 3, no revisions**			
MSE-F		M	.56	.71	.89	.99
MSE-t	S_{dd}	M	.34	.51	.81	.98
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.53	.65	.82	.95
adj. MSE-t	$S_{dd} + 2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.13	.20	.37	.68
$\pi = 0$:						
MSE-F		M	.20	.43	.73	.96
MSE-t	S_{dd}	N	.17	.24	.41	.71
adj. MSE-t	$FBS_{hh}BF'$	N	.52	.63	.77	.92
$\pi > 0$:			**NDGP 3, noise**			
MSE-F		M	.20	.27	.48	.70
MSE-t	S_{dd}	M	.12	.22	.40	.63
adj. MSE-t	$2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.37	.42	.49	.60
adj. MSE-t	$S_{dd} + 2\Pi F\tilde{B}S_{hh}\tilde{B}F'$	N	.07	.08	.11	.19
$\pi = 0$:						
MSE-F		M	.40	.47	.62	.78
MSE-t	S_{dd}	N	.09	.10	.13	.21
adj. MSE-t	$FBS_{hh}BF'$	N	.36	.39	.43	.50

Notes:

1. The DGP is defined in equations (9) and (12). The forecasting models are given in equations (15) and (16).
2. See the notes to Table 4.

Table 6. Correlations of Revisions with Real Time Data

(1965:Q4–2003:Q4 data)

	actual estimate from period:			
revision:	$t+1$	$t+2$	$t+9$	final
GDP growth				
$t+2$ est. $-$ $t+1$ est.	.20	.41	.43	.35
$t+9$ est. $-$ $t+1$ est.	-.11	.02	.34	.23
final est. $-$ $t+1$ est.	-.25	-.18	-.01	.39
$t+9$ est. $-$ $t+2$ est.	-.25	-.23	.14	.06
final est. $-$ $t+2$ est.	-.33	-.33	-.16	.27
HP output gap				
$t+2$ est. $-$ $t+1$ est.	-.87	-.78	-.31	-.26
$t+9$ est. $-$ $t+1$ est.	-.53	-.47	.38	.40
final est. $-$ $t+1$ est.	-.57	-.51	.29	.41
$t+9$ est. $-$ $t+2$ est.	-.36	-.30	.55	.55
final est. $-$ $t+2$ est.	-.41	-.37	.43	.56
Change in GDP inflation				
$t+2$ est. $-$ $t+1$ est.	-.15	.21	.13	.06
$t+9$ est. $-$ $t+1$ est.	-.45	-.31	.26	-.04
final est. $-$ $t+1$ est.	-.64	-.54	-.30	.24
$t+9$ est. $-$ $t+2$ est.	-.40	-.45	.21	-.08
final est. $-$ $t+2$ est.	-.58	-.64	-.37	.21

Notes:

1. For each variable, the table reports correlations between revisions and estimates of GDP growth, the HP output gap, and the change in inflation. The revisions are constructed as a later estimate minus an earlier estimate, specified in the first column of the table. The estimates of actual growth, etc., are taken from various vintages, given in the column headers. “Final” refers to the 2007:Q1 vintage of data from the RTDSM.

Table 7. Results for Non-Nested Model Inflation Forecasts

sample	MSE$_1$	MSE$_2$	MSE$_1$ - MSE$_2$	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	$t(S_{dd})$	$t(\Omega)$
1-step horizon							
actual inflation$_t$ = estimate published in $t+1$							
1970-2003	2.164	2.181	-.017	.171	.179	-.100	-.096
1970-1984	3.791	3.758	.033	.379	.366	.086	.089
1985-2003	.880	.937	-.056	.059	.078	-.964	-.726
actual inflation$_t$ = estimate published in $t+2$							
1970-2003	2.311	2.372	-.061	.165	.174	-.372	-.353
1970-1984	4.033	4.073	-.040	.365	.358	-.111	-.113
1985-2003	.951	1.029	-.078	.061	.078	-1.285	-1.004
actual inflation$_t$ = estimate published in $t+5$							
1970-2003	2.481	2.447	.034	.191	.190	.179	.179
1970-1984	4.489	4.314	.174	.427	.387	.408	.450
1985-2003	.896	.972	-.076	.051	.075	-1.498	-1.018
actual inflation$_t$ = estimate published in $t+13$							
1970-2003	2.252	2.438	-.186	.185	.212	-1.005	-.877
1970-1984	4.196	4.512	-.315	.416	.449	-.759	-.702
1985-2003	.717	.801	-.084	.044	.073	-1.918*	-1.157
4-step horizon							
actual inflation$_t$ = estimate published in $t+1$							
1970-2003	1.563	1.933	-.371	.216	.255	-1.714*	-1.455
1970-1984	2.925	3.591	-.665	.471	.585	-1.413	-1.138
1985-2003	.541	.691	-.150	.079	.106	-1.896*	-1.407
actual inflation$_t$ = estimate published in $t+2$							
1970-2003	1.984	2.361	-.378	.222	.266	-1.703*	-1.420
1970-1984	3.908	4.580	-.673	.484	.631	-1.389	-1.066
1985-2003	.541	.697	-.156	.080	.109	-1.956*	-1.432
actual inflation$_t$ = estimate published in $t+5$							
1970-2003	1.960	2.424	-.464	.257	.314	-1.804*	-1.478
1970-1984	3.866	4.733	-.868	.556	.715	-1.560	-1.214
1985-2003	.532	.692	-.161	.076	.112	-2.122*	-1.432
actual inflation$_t$ = estimate published in $t+13$							
1970-2003	1.994	2.528	-.533	.300	.356	-1.778*	-1.496
1970-1984	3.913	5.005	-1.092	.642	.815	-1.701*	-1.339
1985-2003	.555	.670	-.114	.054	.107	-2.102*	-1.071

Notes:

1. The table compares the accuracy of real-time forecasts of the change in GDP inflation, from equation (17). Model 1 uses x_t = four-quarter GDP growth; Model 2 uses x_t = the output gap, computed with the HP filter. The models are estimated recursively, with the sample beginning in 1961:1+τ-1.
2. The MSEs are defined as annualized percentage points. MSE$_1$ refers to the mean square error of forecasts from the model with GDP growth; MSE$_2$ refers to the mean square error of forecasts from the model with the output gap. 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 variance Ω is defined as $S_{dd} + 2\Pi(FBS_{dh} + FBS_{hh}BF')$. The columns $t(S_{dd})$ and $t(\Omega)$ report t-statistics for the difference in MSEs computed with the variances S_{dd} and Ω, respectively. An * denotes a rejection of the null of equal accuracy at a significance level of 10% or better.

Table 8. Results for Nested Model Inflation Forecasts

sample	MSE$_1$	MSE$_2$	MSE$_1$ - MSE$_2$	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	$t(S_{dd})$	$t(\Omega)$	MSE-F
1-step horizon								
actual inflation$_t$ = estimate published in $t+1$								
1970-2003	2.368	2.164	.203	.176	.099	1.159*	2.055*	12.786*
1970-1984	4.096	3.791	.305	.390	.218	.782*	1.399*	4.829*
1985-2003	1.003	.880	.123	.061	.059	2.026*	2.079*	10.644*
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.359	2.311	.048	.164	.086	.294*	.558	2.841*
1970-1984	3.986	4.033	-.047	.365	.264	-.128	-.176	-.692
1985-2003	1.074	.951	.123	.056	.068	2.191*	1.804*	9.834*
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.565	2.481	.085	.182	.102	.466*	.829	4.646*
1970-1984	4.504	4.489	.016	.406	.278	.039	.057	.211
1985-2003	1.035	.896	.139	.053	.045	2.622*	3.101*	11.813*
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.297	2.252	.045	.162	.090	.278*	.499	2.713*
1970-1984	4.221	4.196	.025	.362	.251	.068	.098	.351
1985-2003	.778	.717	.061	.042	.013	1.467*	4.538*	6.470*
4-step horizon								
actual inflation$_t$ = estimate published in $t+1$								
1970-2003	2.170	1.563	.607	.413	.151	1.469*	4.019*	52.833*
1970-1984	4.262	2.925	1.337	.884	.488	1.511*	2.739*	27.413*
1985-2003	.601	.541	.060	.122	.024	.492*	2.549*	8.435*
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.648	1.984	.664	.493	.169	1.347*	3.927*	45.501*
1970-1984	5.378	3.908	1.471	1.071	.549	1.373*	2.678*	22.579*
1985-2003	.599	.541	.059	.122	.028	.481*	2.064*	8.236
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.600	1.960	.640	.477	.157	1.341*	4.075*	44.368*
1970-1984	5.277	3.866	1.412	1.034	.488	1.365*	2.891*	21.909*
1985-2003	.592	.532	.061	.133	.015	.455*	4.152*	8.664*
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.531	1.994	.536	.468	.134	1.145*	4.018*	36.580*
1970-1984	5.224	3.913	1.311	1.012	.477	1.296*	2.751*	20.109*
1985-2003	.510	.555	-.045	.117	.046	-.382	-.975	-6.137

Notes:

1. The table compares the accuracy of real-time forecasts of the change in GDP inflation, from equations (17) (MSE$_1$) and (18) (MSE$_2$), with x_t measured as four-quarter GDP growth. 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 variance Ω is defined as $2\Pi F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$. The columns $t(S_{dd})$ and $t(\Omega)$ report t-statistics for the difference in MSEs computed with the variances S_{dd} and Ω, respectively. An * denotes a rejection of the null of equal accuracy at a significance level of 10% or better, for the following: $t(S_{dd})$ vs. critical values simulated as in Clark and McCracken (2005); $t(\Omega)$ vs. standard normal critical values; and MSE-F vs. critical values simulated as in Clark and McCracken (2005).