EVALUATING DIRECT MULTI-STEP FORECASTS

Todd Clark and Michael McCracken

Revised: April 2005
(First Version December 2001)

RWP 01-14

Research Division
Federal Reserve Bank of Kansas City

Todd E. Clark is a vice president and economist at the Federal Reserve Bank of Kansas City. Michael W. McCracken is an assistant professor of economics at the University of Missouri-Columbia. Earlier versions of this paper were titled "Evaluating Long--Horizon Forecasts." The authors gratefully acknowledge the helpful comments of Lutz Kilian, David Rapach, Ken West, seminar participants at the Federal Reserve Bank of Kansas City, and participants at the 2001 MEG meetings. McCracken thanks LSU for financial support during work on a substantial portion of this paper. 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 or the Federal Reserve System.

Clark email: todd.e.clark@kc.frb.org
McCracken email: mccrackenm@missouri.edu

Abstract

This paper examines the asymptotic and finite-sample properties of tests of equal forecast accuracy and encompassing applied to direct, multi--step predictions from nested regression models. We first derive the asymptotic distributions of a set of tests of equal forecast accuracy and encompassing, showing that the tests have non-standard distributions that depend on the parameters of the data-generating process. We then conduct a range of Monte Carlo simulations to examine the finite-sample size and power of the tests. In these simulations, our asymptotic approximation yields good finite--sample size and power properties for some, but not all, of the tests; a bootstrap works reasonably well for all tests. The paper concludes with a reexamination of the predictive content of capacity utilization for core inflation.

JEL classification: C53, C12, C52

Keywords: Prediction, long horizon, causality

1 Introduction

Researchers often compare multi–step forecasts from nested linear models to determine whether one variable has predictive content for another. Examples include Estrella and Hardouvelis' (1991) examination of the predictive content of spreads for GDP growth, Mark's (1995) and Kilian's (1999) studies of exchange rate models, and Stock and Watson's (1999, 2003) analyses of output and inflation forecasting models. In such applications, forecasts from the model of interest are compared to forecasts from a benchmark model that is a restricted version of the model of interest. Consequently, the results in studies such as West (1996, 2001) on the asymptotic and finite-sample properties of tests of equal forecast accuracy and encompassing, based on non-nested models, may not apply.[1] Intuitively, with nested models, the null hypothesis that the restrictions imposed in the benchmark model are true implies the population errors of the competing forecasting models are exactly the same. This in turn implies, for example, that the population difference between the competing models' mean square forecast errors is exactly zero with zero variance. As a result, the distribution of a t–statistic for equal MSE may be non–standard. Indeed, Clark and McCracken (2001) and McCracken (2004) show that, for 1-step ahead forecasts from nested models, the distributions of tests for equal forecast accuracy and encompassing can be non-standard.

In many comparisons of multi–step forecasts from nested models, the multi–step predictions are made using horizon–specific, linear models, in which the dependent variable is the multi–step ahead value being forecast. As described in studies such as Clements and Hendry (1996), Schorfheide (2003), Chevillon and Hendry (2004), and Marcellino, Stock, and Watson (2004), an alternative approach is to form multi–step forecasts by iterating forward projections from one–step ahead models. Both methods have pros and cons, reviewed in the aforementioned studies. But one of the key advantages of the direct approach in forecasting is its computational simplicity — an advantage that no doubt helps account for its common usage, in applications such as those listed above.

Motivated by the frequency with which researchers compare direct multi-step predictions from nested linear regression models, this paper examines the asymptotic

[1] As described explicitly in West's (2005) survey, the nesting of the models violates a rank condition required in the asymptotic normality results of West (1996).

and finite-sample properties of tests of equal forecast accuracy and encompassing applied to such forecasts. Our multi–step analysis builds on the one–step analyses of Clark and McCracken (2001) and McCracken (2004). Specifically, for direct, multi–step forecasts from nested models, we first derive the asymptotic distributions of some standard tests of equal forecast accuracy and encompassing and the variants proposed in McCracken (2004) and Clark and McCracken (2001). As in our prior work and other studies such as West (1996, 2001), West and McCracken (1998), Chao, Corradi, and Swanson (2001), Corradi, Swanson, and Olivetti (2001), and Gilbert (2001), the distributions explicitly account for the uncertainty introduced by parameter estimation. In general, the tests have non-standard distributions that depend on the parameters of the data-generating process.

In light of the dependence of the asymptotic distributions on unknown nuisance parameters, in our Monte Carlo analysis of finite–sample size and power and in our empirical application we consider both asymptotic and bootstrap approaches to inference. The asymptotic approach — which could be applied by any researcher — involves estimating the particular second moments of the data that affect the limiting distributions. Our bootstrap procedure is a slightly simplified version of the one Kilian (1999) used in analyzing the predictability of exchange rates. The Monte Carlo results indicate our asymptotic approximation yields good finite–sample size and power properties for some, but not all, of the tests considered; a bootstrap works reasonably well for all tests. Most notably, the asymptotics seem to work well for McCracken's (2004) F–type test of equal MSE, delivering a test with decent size and power properties. But the encompassing test proposed by Clark and McCracken (2001) has superior power (even when based on bootstrap critical values, which generally yield correctly sized tests).[2]

Finally, to illustrate how the tests perform in practical settings, the paper concludes with an examination of capacity utilization's predictive power for core CPI inflation. Cecchetti (1995), Staiger, Stock, and Watson (1997), and Stock and Watson (1999, 2003) are recent examples of studies in the long literature on this basic question. Applying our tests and bootstrap approach to inference to simulated out–of–sample forecasts for 1976-2004, we find that capacity utilization in manufacturing

[2]Clark and McCracken (2005) consider how these out-of-sample tests behave under a broad range of alternatives that include breaks in the causal relationships. See Rossi (2001) and Inoue and Kilian (2004) for further discussion of the power of out-of-sample tests compared to in–sample tests.

has significant predictive power for core inflation.

Although our results apply only to a setup that some might see as restrictive — direct, multi–step (DMS) forecasts from nested linear models — the long list of studies analyzing such forecasts suggests our results should be useful to many researchers. Recent applications considering DMS forecasts from nested linear models include, among others: the studies cited at the beginning of this section; Diebold and Li (2004); Orphanides and van Norden (2004); Rapach and Weber (2004); and Shintani (2004). Of course, a number of other studies, such as Marcellino (2002), Kilian and Taylor (2003), Qi and Wu (2003), have considered DMS forecasts from nested nonlinear models. We leave as an important topic for future research the extension of our asymptotics to allow nonlinear models.[3] Similarly, we leave the extension of our results to iterated multi–step forecasts to future work.[4]

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, appropriate asymptotic critical values can be calculated. Proofs of the asymptotic results are provided in the appendix. Section 4 describes our model–based bootstrap approach and presents Monte Carlo results on the finite–sample performance of the asymptotics and the bootstrap. Section 5 applies our tests to determine whether capacity utilization has predictive power for core inflation. Section 6 concludes.

2 Setup

The sample of observations $\{y_t, x'_{2,t}\}_{t=1}^{T}$ includes a scalar random variable y_t to be predicted, as well as a $(k_1 + k_2 = k \times 1)$ vector of predictors $x_{2,t} = (x'_{1,t}, x'_{22,t})'$. Specifically, for each time t the variable to be predicted is $y_{t+\tau}$, where τ denotes the forecast horizon. The sample is divided into in–sample and out–of–sample portions. The total in–sample observations (on y_t and $x_{2,t}$) span 1 to R. Letting $P - \tau + 1$ denote the number of τ–step ahead predictions, the total out-of-sample observations span $R + \tau$ through $R + P$. The total number of observations in the sample is

[3] Corradi and Swanson (2002) develop an encompassing–type test for comparing one–step ahead forecasts from a pair of nested nonlinear or linear models

[4] Iteration will mean the multi–step forecasts are affected by polynomials in parameter estimation error. In contrast, with DMS forecasts, parameter estimation error enters only linearly.

$R + P = T$.

Forecasts of $y_{t+\tau}$, $t = R, \ldots, T-\tau$, are generated using the two linear models $y_{t+\tau} = x'_{1,t}\beta_1^* + u_{1,t+\tau}$ (model 1) and $y_{t+\tau} = x'_{2,t}\beta_2^* + u_{2,t+\tau}$ (model 2). Under the null hypothesis of equal forecast accuracy or forecast encompassing, model 2 nests model 1 for all t and hence model 2 includes k_2 excess parameters. Then $\beta_2^* = (\beta_1^{*\prime}, 0')'$, and $u_{1,t+\tau} = u_{2,t+\tau} \equiv u_{t+\tau}$ for all t.

Both model 1's and model 2's forecasts are generated recursively using estimated parameters. Under this approach both β_1^* and β_2^* are reestimated with added data as forecasting moves forward through time: for $t = R, \ldots, T-\tau$, model i's (i = 1,2) prediction of $y_{t+\tau}$ is created using the parameter estimate $\hat{\beta}_{i,t}$ based on data through period t.[5] Models 1 and 2 yield two sequences of $P - \tau + 1$ forecast errors, denoted $\hat{u}_{1,t+\tau} = y_{t+\tau} - x'_{1,t}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+\tau} = y_{t+\tau} - x'_{2,t}\hat{\beta}_{2,t}$, respectively. Asymptotic results for forecasts based on the rolling and fixed schemes, described in West and McCracken (1998), are provided in Clark and McCracken (2004).

Finally, the asymptotic results presented below use the following additional notation. Let $h_{i,t+\tau}(\beta_i) = (y_{t+\tau} - x'_{i,t}\beta_i)x_{i,t}$, $h_{i,t+\tau} = h_{i,t+\tau}(\beta_i^*)$, $q_{i,t} = x_{i,t}x'_{i,t}$, $B_i = (Eq_{i,t})^{-1}$and $Eu_{t+\tau}^2 = \sigma^2$. For $H_2(t)$ defined in Assumption 1, J the selection matrix $(I_{k_1 \times k_1}, 0_{k_1 \times k_2})'$, and a $(k_2 \times k)$ matrix $\tilde{A}$ satisfying $\tilde{A}'\tilde{A} = B_2^{-1/2}(-J'B_1J + B_2)B_2^{-1/2}$, let $\tilde{h}_{t+\tau} = \sigma^{-1}\tilde{A}B_2^{1/2}h_{2,t+\tau}$ and $\tilde{H}_2(t) = \sigma^{-1}\tilde{A}B_2^{1/2}H_2(t)$. If we define $\Gamma_{\tilde{h}\tilde{h}}(i) = E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-i}$, then $S_{\tilde{h}\tilde{h}} = \Gamma_{\tilde{h}\tilde{h}}(0) + \sum_{i=1}^{\tau-1}(\Gamma_{\tilde{h}\tilde{h}}(i) + \Gamma'_{\tilde{h}\tilde{h}}(i))$. Let $W(\omega)$ denote a $k_2 \times 1$ vector standard Brownian motion. For the sequence $U_{t+\tau}$ defined in Assumption 2, $U(t)$ is defined analogously to $H(t)$ in Assumption 1.

Given the definitions and forecasting scheme described above, the following assumptions are used to derive the limiting distributions in Theorems 3.1-3.4. The assumptions are intended to be only sufficient, not necessary and sufficient.

(A1) The parameter estimates $\hat{\beta}_{i,t}$, $i = 1, 2$, $t = R, ..., T-\tau$, satisfy $\hat{\beta}_{i,t} - \beta_i^* = B_i(t)H_i(t)$ where $B_i(t)H_i(t) = (t^{-1}\sum_{j=1}^{t-\tau} q_{i,j})^{-1}(t^{-1}\sum_{j=1}^{t-\tau} h_{i,j+\tau})$.

(A2) (a) $U_{t+\tau} = [u_{t+\tau}, x'_{2,t} - Ex'_{2,t}, h'_{2,t+\tau}]'$ is covariance stationary, (b) $EU_{t+\tau} = 0$, (c) $Eq_{2,t} < \infty$ and is positive definite, (d) For some $r > 8$, $U_{t+\tau}$ is uniformly L^r

[5] For the purposes of forecasting, in our setup the largest number of observations used to estimate each model's parameters is $T - 2\tau$. With the dependent variable $y_{t+\tau}$, τ observations are lost in forming the dependent variable and another τ observations are needed for forming the first τ-period out-of-sample forecast.

bounded, (e) For some $r > d > 2$, $U_{t+\tau}$ is strong mixing with coefficients of size $-rd/(r-d)$, (f) With $\tilde{U}_{t+\tau}$ denoting the vector of nonredundant elements of $U_{t+\tau}$, $lim_{T\to\infty} T^{-1} E(\sum_{s=1}^{T-\tau} \tilde{U}_{s+\tau})(\sum_{s=1}^{T-\tau} \tilde{U}_{s+\tau})' = \Omega < \infty$ is positive definite.

(A3) (a) Let $K(x)$ be a continuous kernel such that for all real scalars x, $|K(x)| \leq 1$, $K(x) = K(-x)$ and $K(0) = 1$, (b) For some bandwidth M and constant $i \in (0, 0.5)$, $M = O(P^i)$, (c) For all $j > \tau - 1$, $Eh_{2,t+\tau} h'_{2,t+\tau-j} = 0$, (d) The number of covariance terms $\bar{j}$, used to estimate the long–run covariances S_{cc} and S_{dd} defined in Section 3.1, satisfies $\tau - 1 \leq \bar{j} < \infty$.

(A4) $\lim_{R,P\to\infty} P/R = \pi \in (0, \infty)$; define $\lambda = (1+\pi)^{-1}$.

(A4′) $\lim_{R,P\to\infty} P/R = 0$; define $\lambda = 1$.

The assumptions provided here are broadly similar to those provided in Clark and McCracken (2001) and McCracken (2004). We restrict attention to forecasts generated using parameters estimated by OLS (Assumption 1) and we do not allow for processes with either unit roots or time trends (Assumption 2).[6] We provide asymptotic results for situations in which the in-sample and out-of-sample sizes R and P are of the same order (Assumption 4) as well as when the in-sample size R is large relative to the out-of-sample size P (Assumption 4′).

Assumption 3 is necessitated by the serial correlation in the multi-step (τ-step) forecast errors — errors from even well-specified models exhibit serial correlation, of an MA($\tau - 1$) form. Typically, researchers constructing a t-statistic utilizing the squares of these errors account for serial correlation of at least order $\tau - 1$ in forming the necessary standard error estimates. Meese and Rogoff (1988), Groen (1999), and Kilian and Taylor (2003), among other applications to forecasts from nested

[6] Our assumptions do, however, allow y_t and $x_{2,t}$ to be stationary differences of trending variables. As to other technical aspects of Assumption 2, (a) and (c) together ensure that in large samples, sample averages of the outer product of the predictors will be invertible and hence the least squares estimate will be well defined. Part (d) enables the use of Markov inequalities when showing certain terms are asymptotically negligible. Along with (d), (e) and (f) allow us to use results in Hansen (1992) and Davidson (1994) regarding the weak convergence of partial sums to Brownian motion and that functionals of these partial sums converge in distribution to stochastic integrals. To ensure the variance matrix non-singularity required for weak convergence, in (f) we eliminate the possibility that $\tilde{U}_{t+\tau}$ has elements that are identical by defining it to include only the nonredundant elements of $U_{t+\tau}$. For example, if the unrestricted forecasting model includes a constant, $U_{t+\tau}$ will include $u_{t+\tau}$ twice, once directly and again as the first element of $h_{2,t+\tau}$.

models, use kernel-based methods to estimate the relevant long-run covariance.[7] We therefore impose conditions sufficient to cover applied practices. Parts (a) and (b) are not particularly controversial. Part (c), however, imposes the restriction that the orthogonality conditions used to identify the parameters form a moving average of finite order $\tau - 1$, while part (d) imposes the restriction that this fact is taken into account when constructing the MSE-T and ENC-T statistics discussed in Section 3.[8] Although Assumption 3 and our theoretical results admit a range of kernel and bandwidth approaches, in our Monte Carlo experiments and empirical application we compute the variances required by the MSE-T and ENC-T t–statistics (for $\tau > 1$) using the Newey and West (1987) estimator with a lag length of $1.5 * \tau$.

The above assumptions differ importantly from those underlying our previous work, in that we do not require the forecast errors to form a conditionally homoskedastic martingale difference sequence. Rather, we allow for conditional heteroskedasticity and the effects of serial correlation induced by forecast horizons greater than one period. In contrast, our prior work considered only conditionally homoskedastic, serially uncorrelated, one–step ahead forecast errors. Nevertheless, our assumptions remain strong enough for us to use Hansen's (1992) and Davidson's (1994) theoretical results regarding weak convergence of partial sums to Brownian motion and averages of these partial sums to stochastic integrals of Brownian motion. As we will see below, the null limiting distributions bear a strong resemblance to those in Clark and McCracken (2001) and McCracken (2004), but depend upon unknown nuisance parameters.

3 Tests and Asymptotic Distributions

We consider a total of four forecast–based tests, two tests of equal forecast accuracy and two tests for forecast encompassing. In particular, we consider the t–statistic

[7] For similar uses of kernel–based methods in analyses of non–nested forecasts, see, for example, Diebold and Mariano (1995) and West (1996).

[8] We have bounded the numbers of covariances used to construct $\hat{S}_{dd}$ and $\hat{S}_{cc}$ in order to be able to derive asymptotic results for the MSE-T and ENC-T tests. Technically, without any bounds on the bandwidth, we would have to find the limiting behavior of a kernel–weighted infinite sum of individually $o_p(1)$ elements. Because it is unclear how this would be accomplished, for tractability we restrict the number of autocovariances for which $Eh_{2,t+\tau}h'_{2,t+\tau-j} \neq 0$ to be finite and take this into account when constructing both $\hat{S}_{cc}$ and $\hat{S}_{dd}$.

for equal MSE developed by Diebold and Mariano (1995) and West (1996) and the F–statistic proposed by McCracken (2004). We also consider the t–statistic for encompassing developed in Harvey, Leybourne, and Newbold (1998) and West (2001) and the variant proposed by Clark and McCracken (2001). In preliminary Monte Carlo results, regression–based variants of the t–statistics for equal MSE and forecast encompassing, proposed respectively by Granger and Newbold (1977) and Ericsson (1992), performed similarly to the versions considered below. As a result, in the interest of brevity, we leave these regression–based tests out of the analysis below.

3.1 t–type tests: MSE-T and ENC-T

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal MSE based upon the sequence of loss differentials $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$. If we define $\text{MSE}_i = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{u}^2_{i,t+\tau}$ $(i=1,2)$, $\bar{d} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau} = \text{MSE}_1 - \text{MSE}_2$, $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}(\hat{d}_{t+\tau}-\bar{d})(\hat{d}_{t+\tau-j}-\bar{d})$, $\hat{\Gamma}_{dd}(-j) = \hat{\Gamma}_{dd}(j)$, and $\hat{S}_{dd} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{dd}(j)$, 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 $x_{22,t}$ has no predictive power for $y_{t+\tau}$, the population difference in MSEs will equal 0. Under the alternative that $x_{22,t}$ has predictive power, the population difference in MSEs will be positive ($\text{MSE}_2 < \text{MSE}_1$). As a result, the MSE-T test and the other equal MSE test described below are one–sided to the right.

Drawing on the methodology of Diebold and Mariano (1995), Harvey, Leybourne, and Newbold (1998) propose a test of encompassing that uses a t–statistic for the covariance between $u_{1,t+\tau}$ and $u_{1,t+\tau} - u_{2,t+\tau}$. If we define $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$, $\bar{c} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}$, $\hat{\Gamma}_{cc}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}(\hat{c}_{t+\tau}-\bar{c})(\hat{c}_{t+\tau-j}-\bar{c})$, $\hat{\Gamma}_{cc}(-j) = \hat{\Gamma}_{cc}(j)$, and $\hat{S}_{cc} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{cc}(j)$, the statistic takes the form

$$\text{ENC-T} = (P-\tau+1)^{1/2} \times \frac{\bar{c}}{\sqrt{\hat{S}_{cc}}}. \tag{2}$$

Under the null that $x_{22,t}$ has no predictive power for $y_{t+\tau}$, the population covariance between $u_{1,t+\tau}$ and $u_{1,t+\tau} - u_{2,t+\tau}$ will equal 0 (the population forecast errors of the models will be exactly the same). Under the alternative that $x_{22,t}$ does have predictive power, the covariance will be positive. To see why, consider the forecast combination

regression $y_{t+\tau} = (1-\alpha)f_{1,t+\tau} + \alpha f_{2,t+\tau} + \text{error}$, where f_1 and f_2 denote forecasts from the restricted and unrestricted models, respectively.[9] Subtracting $f_{1,t}$ from both sides, and making the substitution $u_{1,t+\tau} - u_{2,t+\tau} = f_{2,t+\tau} - f_{1,t+\tau}$, yields the encompassing regression $u_{1,t+\tau} = \alpha(u_{1,t+\tau} - u_{2,t+\tau}) + \text{error}$. If $x_{22,t}$ does have predictive power, such that model 2 is true, the population combination coefficient α equals 1. As a result, the covariance between $u_{1,t+\tau}$ and $(u_{1,t+\tau} - u_{2,t+\tau})$ will be positive. Consequently, the ENC-T test and the other forecast encompassing test described below are one–sided to the right.

While West (1996) proves directly that the MSE-T statistic can be asymptotically standard normal when applied to non–nested forecasts and West's results suffice to establish the same for the ENC-T statistic, this is not the case when the models are nested. In particular, the results in West require that under the null, the population–level long run variances of $\hat{d}_{t+\tau}$ and $\hat{c}_{t+\tau}$ be positive. This requirement is violated with nested models. Intuitively, with nested models, the null hypothesis that the restrictions imposed in the benchmark model are true implies the population errors of the competing forecasting models are exactly the same. As a result, in population $d_{t+\tau} = 0$ and $c_{t+\tau} = 0$ for all t, which makes the corresponding variances also equal to 0. Because the sample analogues (for example, $\bar{d}$ and its variance) converge to zero at the same rate, the test statistics have non–degenerate null distributions, but they are non–standard.

Specifically, McCracken (2004) shows that, for 1–step ahead forecasts from well-specified nested models, the MSE-T test statistic converges in distribution to a function of stochastic integrals of quadratics of Brownian motion, with a limiting distribution that depends on the sample split parameter π and the number of exclusion restrictions k_2 but does not depend upon any unknown nuisance parameters. Under the same conditions, Clark and McCracken (2001) show that the ENC-T test statistic converges to the same type of distribution. With direct multi-step forecasts, however, the limiting distributions are affected by unknown nuisance parameters. (Note that, for these particular asymptotic results, we present the ENC-T theorem before the MSE-T theorem because, analytically, it is easiest to first establish the ENC-T results and then use those in deriving the MSE-T asymptotics.)

[9] This basic logic is laid out in Harvey, Leybourne, and Newbold (1998), in the context of non–nested models.

Theorem 3.1. (a) Let Assumptions 1-4 hold. For ENC-T defined in (2), ENC-T $\rightarrow_d \Gamma_1/\Gamma_3^{1/2}$, where $\Gamma_1 = \int_\lambda^1 s^{-1}W(\omega)'S_{\tilde{h}\tilde{h}}dW(\omega)$ and $\Gamma_3 = \int_\lambda^1 s^{-2}W(\omega)'S_{\tilde{h}\tilde{h}}^2W(\omega)d\omega$. (b) Let Assumptions 1-3 and 4′ hold and let V_0 and V_1 denote $(k_2 \times 1)$ independent standard normal vectors. ENC-T $\rightarrow_d V_0'S_{\tilde{h}\tilde{h}}V_1/[V_0'S_{\tilde{h}\tilde{h}}^2V_1]^{1/2} \sim N(0,1)$.

Theorem 3.2. (a) Let Assumptions 1-4 hold and define $\Gamma_2 = \int_\lambda^1 s^{-2}W(\omega)'S_{\tilde{h}\tilde{h}}W(\omega)d\omega$. For MSE-T defined in (1) and Γ_1 and Γ_3 defined in Theorem 3.1, MSE-T $\rightarrow_d (\Gamma_1 - (0.5)\Gamma_2)/\Gamma_3^{1/2}$. (b) Let Assumptions 1-3 and 4′ hold. MSE-T – ENC-T $= o_p(1)$.

The results in Theorems 3.1 (a) and 3.2 (a) bear a strong resemblance to those discussed in Clark and McCracken (2001) and McCracken (2004), but with one major distinction: the limiting null distributions generally depend upon the unknown nuisance parameter $S_{\tilde{h}\tilde{h}}$ that in turn depends upon the second moments of the forecast errors $u_{t+\tau}$, the regressors $x_{2,t}$, and the orthogonality conditions $h_{2,t+\tau}$. Algebraically, this dependence arises because, in the presence of conditional heteroskedasticity or serial correlation in the forecast errors, an information matrix-type equality fails: the expected outer product of the predictors is no longer proportional to the long run variance of $h_{2,t+\tau}$ with constant of proportionality $\sigma^2 = Eu_{2,t+\tau}^2$. Similarly, in the context of likelihood-ratio statistics, Vuong (1989, Theorem 3.3) shows that the limiting distribution of the likelihood ratio statistic has a representation as a mixture of independent $\chi_{(1)}^2$ variates (in contrast to our integrals of weighted quadratics of Brownian motion). This distribution is free of nuisance parameters when the information matrix equality holds but in general does depend upon such nuisance parameters.

In Theorems 3.1 and 3.2 there are, however, special cases for which the dependence on $S_{\tilde{h}\tilde{h}}$ is asymptotically irrelevant. When $k_2 = 1$ the now scalar $S_{\tilde{h}\tilde{h}}$ can be factored out of both the numerator and denominator and hence cancels. Also, in the perhaps unlikely scenario in which each of the eigenvalues of $S_{\tilde{h}\tilde{h}}$ are identical, one can show that the limiting distributions no longer depend upon the value of $S_{\tilde{h}\tilde{h}}$. If either of these special cases hold we obtain McCracken's (2004) results for MSE-T and Clark and McCracken's (2001) results for ENC-T and thus are able to utilize the estimated asymptotic critical values provided in those papers to conduct inference. In general, though, the distributions do depend upon $S_{\tilde{h}\tilde{h}}$ and hence those critical

values are no longer relevant. Instead, as described below, we consider estimating the asymptotically valid critical values both by simulating the asymptotic distribution implied by a consistent estimate of $S_{\tilde{h}\tilde{h}}$ and by bootstrapping the distribution.

Note also that, in line with the results of Clark and McCracken (2001), for case (b) we find that the MSE-T and ENC-T statistics are asymptotically equivalent under the null. They are also asymptotically standard normal. On a practical level this implies that for instances in which the number of out-of-sample observations P is small relative to the number of out-of-sample observations R, we should expect these two test statistics to behave similarly, at least under the null. Moreover, inference is straightforward since appropriate critical values are readily obtained.

In light of the standard normality that applies when $\pi = \lim_{R,P\to\infty} P/R = 0$, a natural question is, in practice, how small must P be relative to R for standard normal critical values to be reliably used? The answer is that P/R has to be considerably smaller than it is in most studies. Simulations for one–step ahead forecasts in Clark and McCracken (2001) suggest that standard normal critical values can reasonably be used for the MSE-T and ENC-T tests when P/R is about .10.[10] Even when P/R is just .20, our asymptotics are more reliable than a standard normal approximation. We corroborate this rough cutoff of .10 in the simulations reported in section 4. In most historical forecast applications, though, P/R seems to be safely above .10. Select examples from the nested DMS literature include: Estrella and Hardouvelis (1991), $P/R \approx 1.25$; Mark (1995), 1.1; Stock and Watson (2003), .6 and 1.3; Diebold and Li (2004), .8; and Shintani (2004), .6. More generally, suppose we have a forecast sample of five years — a sample that would be quite short by the standards of the literature that motivates our work. For P/R to be .10, we would need another 50 years of data for initial model estimation. Few data samples span 55 years, due to wars, methodological changes in measurement, etc. Accordingly, the non–normality of the MSE-T and ENC-T tests associated with $\pi > 0$ is likely to be very relevant.

3.2 F–type tests: MSE-F and ENC-F

Motivated by (i) the degeneracy of the long-run variance of $d_{t+\tau}$ and (ii) the functional form of the standard in-sample F-test, McCracken (2004) develops an out–of–sample

[10]Similarly, in the context of non–nested models, West's (1996) simulations indicate that P/R needs to be about .10 for parameter estimation error to become irrelevant.

F–type test of equal MSE, given by

$$\text{MSE-F} = (P - \tau + 1) \times \frac{\text{MSE}_1 - \text{MSE}_2}{\text{MSE}_2} = (P - \tau + 1) \times \frac{\bar{d}}{\text{MSE}_2}. \tag{3}$$

Similarly motivated by issues relating to the long-run variance of $c_{t+\tau}$, Clark and McCracken (2001) propose a variant of the ENC-T statistic in which the covariance between $\hat{u}_{1,t+\tau}$ and $\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau}$ is scaled by the estimated variance of one of the forecast errors (for consistency with the other tests considered, here we replace Clark and McCracken's original label "ENC-NEW" with "ENC-F"):

$$\text{ENC-F} = (P - \tau + 1) \times \frac{\bar{c}}{\text{MSE}_2}. \tag{4}$$

Like the t-type tests, the limiting distributions of these F–type tests are non–standard when the forecasts are nested under the null. Clark and McCracken (2001) and McCracken (2004) show that, for one–step ahead forecasts from well-specified nested models, the MSE-F and ENC-F statistics converge in distribution to functions of stochastic integrals of quadratics of Brownian motion, with limiting distributions that depend on the sample split parameter π and the number of exclusion restrictions k_2, but not any unknown nuisance parameters. Again, though, this result is specific to one-step ahead forecasts from well-specified models. For direct multi-step forecasts the limiting distributions are affected by unknown nuisance parameters.

Theorem 3.3. Let Assumptions 1, 2 and 4 hold. For MSE-F defined in (3) and Γ_1 and Γ_2 defined in Theorems 3.1 and 3.2, respectively, MSE-F $\rightarrow_d 2\Gamma_1 - \Gamma_2$. (b) Let Assumptions 1, 2 and 4′ hold. For the $(k_2 \times 1)$ independent standard normal vectors V_0 and V_1 defined in Theorem 3.1(b), $(R/P)^{1/2}$ MSE-F $\rightarrow_d 2V_0' S_{\tilde{h}\tilde{h}} V_1$.

Theorem 3.4. (a) Let Assumptions 1, 2 and 4 hold. For ENC-F defined in (4) and Γ_1 defined in Theorem 3.1, ENC-F $\rightarrow_d \Gamma_1$. (b) Let Assumptions 1, 2 and 4′ hold. $2(R/P)^{1/2}$ ENC-F $- (R/P)^{1/2}$ MSE-F $= o_p(1)$.

Theorems 3.3 (a) and 3.4 (a) show that, as with the t-type tests presented above, if $\pi > 0$ the limiting distributions of the MSE-F and ENC-F tests are neither normal nor chi-square when the forecasts are nested under the null. And again, the limiting

distributions are free of nuisance parameters in only very special cases. In particular, the distributions here are free of nuisance parameters only if $S_{\tilde{h}\tilde{h}} = I$. If this is the case — if, for example, $\tau = 1$ and the forecast errors are conditionally homoskedastic — the MSE-F representation in Theorem 3.3 simplifies to McCracken's (2004) and the ENC-F result in Theorem 3.4 simplifies to Clark and McCracken's (2001), which would allow their estimated asymptotic critical values to be used in conducting inference. Since, in general, that is not the case we again consider both simulating the asymptotic distribution implied by a consistent estimate of $S_{\tilde{h}\tilde{h}}$ and using bootstrap methods to estimate the asymptotically valid critical values. Note also that, as indicated in Theorem 3.3 (b) and Theorem 3.4 (b), when the number of out-of-sample observations P is small relative to the number of in-sample observations R, the MSE-F and ENC-F statistics require re-scaling in order to obtain non-degenerate limiting distributions, even though the distributions for the t-type tests do not.

3.3 Constructing Asymptotic Critical Values

As indicated above, the asymptotic distributions of the forecast tests differ from the conditionally homoskedastic, one–step ahead case considered in Clark and McCracken (2001) and McCracken (2004) in that the quadratics in Brownian motion are weighted by the long–run variance $S_{\tilde{h}\tilde{h}}$. Accordingly, appropriate critical values can be constructed — for any application, by any researcher — using a consistent estimate of this variance matrix and the numerical methods of Clark and McCracken and McCracken. For each data set, we calculate asymptotic critical values as follows, separately for each forecast horizon. Note that, to make the estimate of $S_{\tilde{h}\tilde{h}}$ as precise as possible, we use the full sample of available $(R + P)$ observations in estimating the moments that enter the variance. In the case of conditionally homoskedastic, one–step ahead forecast errors (for which $S_{\tilde{h}\tilde{h}} = I$), the resulting critical values would be exactly the same as those of Clark and McCracken and McCracken.[11]

1. After fitting the restricted forecasting model (in order to impose the null) to the full sample of available data and saving the residuals $\hat{u}_{t+\tau}$, estimate $\hat{S}_{hh}$ = long run variance$(X_{t,2}\hat{u}_{t+\tau})$ with the Newey and West (1987) estimator and a bandwidth of $1.5 * \tau$ for $\tau > 1$ and 0 for $\tau = 1$.

[11] As noted above, when $k_2 = 1$, the critical values of the MSE-T and ENC-T tests are the same regardless of the presence of conditional heteroskedasticity or serial correlation.

2. Using estimates $\hat{B}_i = (\sum_{t=1}^{T=R+P} x_{i,t}x'_{i,t})^{-1}$, $\hat{S}_{12} = (\sum_{t=1}^{T=R+P} x_{22,t}x'_{1,t})$, $\hat{\sigma}^2$ = residual variance from the model estimated in step 1, $\hat{D} = \hat{B}_2^{-1} - \hat{S}_{12}\hat{B}_1\hat{S}'_{12}$, and $\hat{D}^{.5}$ = the Cholesky decomposition of $\hat{D}$, form

$$\hat{S}_{\tilde{h}\tilde{h}} = \hat{\sigma}^{-2} \begin{pmatrix} 0_{k_2 \times k_1} & \hat{D}^{.5} \end{pmatrix} \hat{B}_2 \hat{S}_{hh} \hat{B}_2 \begin{pmatrix} 0_{k_1 \times k_2} \\ \hat{D}^{.5} \end{pmatrix}. \tag{5}$$

3. Compute the eigenvalues of $\hat{S}_{\tilde{h}\tilde{h}}$.

4. Construct 5000 independent draws from the asymptotic distribution of each test statistic, given k_2 and $\widehat{\pi} = P/R$. In generating these draws, the necessary k_2 Brownian motions are simulated as random walks each using an independent sequence of 10,000 i.i.d. $N(0, 10,000^{-.5})$ increments. The integrals are emulated by summing the relevant weighted quadratics of the random walks, using the eigen values of $\hat{S}_{\tilde{h}\tilde{h}}$ as weights. The 10% critical value is calculated as the 90% percentile of the resulting statistics.

4 Monte Carlo Evidence

We use simulations of bivariate DGPs based on common empirical applications to evaluate the finite sample properties of the above tests for equal forecast MSE and encompassing. In these simulations, the restricted forecasting model is a simple autoregression; the unrestricted model adds lags of some other variable of interest. Under the null hypothesis, the additional variables incorporated in the unrestricted model have no predictive content. Because the dependence of the limiting distributions of the test statistics on unknown nuisance parameters rules out simply looking up appropriate critical values in a table, we consider two possible approaches, one based on asymptotics and the other a simple bootstrap. The asymptotic approach, described in section 3.3, involves estimating the long–run variance matrix $S_{\tilde{h}\tilde{h}}$ that enters the limiting distribution of each test statistic and simulating Brownian motions. Because these asymptotic critical values require non-trivial calculations, some researchers might find simple bootstrap methods, used in such studies as Mark (1995), Kilian (1999), and Stock and Watson (2003), to be a natural alternative. Of course, the bootstrap might also be favored for its prospects of better approximating the small sample distributions of the tests.

We proceed by first describing our Monte Carlo framework and bootstrap procedure. We then present results on the size and power of the forecast–based tests. Note that, while the analytical results above defined the predictand in the general form

$y_{t+\tau}$ to simplify notation, in this section we follow common practice in DMS prediction applications and explicitly define the variable of interest as a τ-period change of the form $Y_{t+\tau} - Y_t$ or, for $\tau = 1$, ΔY_{t+1}. The forecasting models relate $Y_{t+\tau} - Y_t$ to lags of the change in Y and a potentially Granger–causal variable denoted x.

4.1 Monte Carlo Design

For two different DGPs (two for both size and power), we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP. The initial observations necessitated by the lag structure of each DGP are generated with draws from the unconditional normal distribution implied by the DGP. We consider results for a variety of forecast horizons: $\tau = 1, 2, 4, 8$, and 12 periods. Similarly, with quarterly data primarily in mind, we also consider a range of sample sizes (R, P), ranging from 60,40 to 60,120 to 200,40.

4.1.1 Size design

The first DGP (DGP-1) is motivated by the literature on the predictive content of spreads for output growth — examples include Estrella and Hardouvelis (1991) and Estrella, Rodrigues, and Schich (2003). In this case, Y is the log of real GDP (scaled by 400 to make ΔY an annualized percentage change) and x is the 10-year government debt yield less the 1-year government debt rate. The DGP is parameterized using model estimates based on quarterly 1959:1-2004:3 data:

$$\begin{aligned} \Delta Y_t &= .242\Delta Y_{t-1} + .149\Delta Y_{t-2} + u_t \qquad (6) \\ x_t &= -.029\Delta Y_{t-1} - .022\Delta Y_{t-2} + 1.141x_{t-1} - .595x_{t-2} + .707x_{t-3} \\ &\quad - .477x_{t-4} + .435x_{t-5} - .428x_{t-6} + .129x_{t-7} + v_t \\ &\quad \text{var}\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 10.265 & \\ -.218 & .159 \end{pmatrix}. \end{aligned}$$

Note that while constants were included in the equations fit to historical data, for simplicity the intercepts have been dropped from the DGP, without any consequence for the results. The lag orders were determined on an equation–by–equation basis using the AIC.[12]

[12] In some instances, we dropped sets of terms in the AIC–determined model that were clearly insignificant. In DGP-1's x equation, for example, the optimal lag length (imposing the same lag length on the both variables in the x equation) was 7. But lags 3-7 of ΔY were insignificant and therefore dropped.

The second DGP is motivated by the inflation forecasting work of Cecchetti (1995), Staiger, Stock, and Watson (1997), and Stock and Watson (1999, 2003), which relates inflation to measures of real activity. Following Stock and Watson (1999, 2003), we presume a unit root in inflation, and make Y the log difference of the quarterly core CPI (scaled by 400 to make Y an annualized percentage change), so that ΔY is the change in quarterly inflation. We specify x as the rate of capacity utilization in manufacturing. The DGP is parameterized using model estimates based on quarterly 1957:1-2004:3 data:

$$\begin{aligned} \Delta Y_t &= -.316\Delta Y_{t-1} - .214\Delta Y_{t-2} + u_t \\ x_t &= -.193\Delta Y_{t-1} - .242\Delta Y_{t-2} - .240\Delta Y_{t-3} - .119\Delta Y_{t-4} \\ &\quad + 1.427x_{t-1} - .595x_{t-2} + .294x_{t-3} - .174x_{t-4} + v_t \\ &\quad \text{var}\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 1.792 & \\ .244 & 1.463 \end{pmatrix}. \end{aligned} \tag{7}$$

4.1.2 Power design

In our power experiments, the x_t equation in each DGP is the same as in the size experiments. Only the ΔY_t equations and the error variance–covariance matrices differ. The equation for ΔY_t in DGP-1, based on an estimated regression of GDP growth on lags of growth and the spread, takes the form

$$\begin{aligned} \Delta Y_t &= .197\Delta Y_{t-1} + .202\Delta Y_{t-2} - .080x_{t-1} + 2.233x_{t-2} - 1.518x_{t-3} + u_t \\ &\quad \text{var}\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 9.067 & \\ -.218 & .159 \end{pmatrix}. \end{aligned} \tag{8}$$

In the second DGP, based on an estimated regression of the change in core inflation on lags of itself and capacity utilization, the ΔY_t equation takes the form

$$\begin{aligned} \Delta Y_t &= -.419\Delta Y_{t-1} - .258\Delta Y_{t-2} \\ &\quad +.331x_{t-1} - .423x_{t-2} + .309x_{t-3} - .139x_{t-4} + u_t \\ &\quad \text{var}\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 1.517 & \\ .244 & 1.463 \end{pmatrix}. \end{aligned} \tag{9}$$

4.1.3 Forecast evaluation

Each Monte Carlo simulation involves first estimating restricted and unrestricted DMS forecasting models. We use models of the form common in the literatures from which we take the applications. For DGP-1, we follow the work of Estrella

and Hardouvelis (1991) and Stock and Watson (2003) on the GDP growth–spread relationship and suppose a restricted model that includes lags of ΔY as predictors and an unrestricted model that adds lags of x to the baseline specification:

$$Y_{t+\tau} - Y_t = \alpha + \sum_{l=0}^{L-1} \gamma_l \Delta Y_{t-l} + u_{1,t+\tau} \tag{10}$$

$$Y_{t+\tau} - Y_t = \alpha + \sum_{l=0}^{L-1} \gamma_l \Delta Y_{t-l} + \sum_{m=0}^{M-1} \beta_m x_{t-m} + u_{2,t+\tau}. \tag{11}$$

The forecasting equations for DGP-2 take a similar form, except that, as in Stock and Watson (1999, 2003), the dependent variable is the difference between a τ-period inflation rate and the lagged quarterly inflation rate (rather than simply a difference between quarterly inflation rates):

$$Y_{t+\tau}^{(\tau)} - Y_t = \alpha + \sum_{l=0}^{L-1} \gamma_l \Delta Y_{t-l} + u_{1,t+\tau} \tag{12}$$

$$Y_{t+\tau}^{(\tau)} - Y_t = \alpha + \sum_{l=0}^{L-1} \gamma_l \Delta Y_{t-l} + \sum_{m=0}^{M-1} \beta_m x_{t-m} + u_{2,t+\tau}, \tag{13}$$

where $Y_{t+\tau}^{(\tau)} = (1/\tau) \sum_{s=1}^{\tau} Y_{t+s}$. $Y_{t+\tau}^{(\tau)}$ corresponds to the average annual rate of price increase from period t to $t+\tau$. Note that the sets of regressors in (10) and (12) correspond to $x_{1,t}$ in our theoretical setup, while the sets of regressors in (11) and (13) correspond to $x_{2,t}$, with x_t and its lags representing $x_{22,t}$.

For each artificial data set, we follow the precedent of such studies as Stock and Watson (2003) and Granger and Jeon (2004) and use data–determined lag orders. Specifically, the lags in the forecasting models are determined by applying the AIC to models estimated with just the first R observations of the sample (we use just the first R observations rather than the whole sample to avoid the type of overfitting that Clark (2004) shows can lead to spurious forecast inference). In the unrestricted forecasting model, we allow the lag orders of ΔY and x to differ, from a range of 0 to 8 for ΔY and 1 to 8 for x. The restricted forecasting model uses the same lag order for ΔY that the unrestricted model does. The lag lengths are allowed to differ across forecast horizons. Of course, the data dependence of the lag orders means that, as is the case in practical applications, the estimated forecasting models may be misspecified.

Following the model estimation, the MSE-T, ENC-T, MSE-F, and ENC-F statistics are formed. The heteroskedasticity and autocorrelation–consistent (HAC) vari-

ances required by the MSE-T and ENC-T t–statistics are calculated using the Newey and West (1987) estimator with a lag length of $1.5 * \tau$ for $\tau > 1$ and 0 for $\tau = 1$. The statistics computed with the Monte Carlo data from a given draw represent the "sample" statistics. We compare these sample statistics against both asymptotic and bootstrap critical values. Based on 1000 Monte Carlo draws, we report the percentage of Monte Carlo trials in which the null of no predictive content is rejected — the percentage of trials in which the sample test statistics exceed the critical values (reporting separate results for asymptotic and bootstrap critical values). In the reported results, the tests are compared against 10% critical values, so that the nominal size of the tests is 10%. Using 5% critical values yields similar findings.

To give a sense of how using standard normal critical values may affect inference, for some tests we also report size and power results based on simply comparing the "sample" test statistics from our 1000 Monte Carlo draws against the standard normal distribution. The set of tests for which we report these results are those researchers sometimes compare against standard critical values: MSE-T and ENC-T . The limiting distributions of these test statistics are standard normal if the forecasting models are non–nested, but the distributions are generally non–standard when the forecasting models are nested.

4.2 Bootstrap Algorithm

Following Berkowitz and Kilian's (2000) recommendations for time series data, our bootstrap algorithm — based on Kilian's (1999) — relies on parametric methods. Vector autoregressive equations for ΔY_t and x_t — restricted to impose the null that x has no predictive power for Y — are estimated by OLS using the full sample of observations, with the residuals stored for sampling. Note that the DGP equation for ΔY takes exactly the same form as the restricted forecasting model for $\tau = 1$ (but estimated with all available data). In the case of the x equation, the lag orders for ΔY and x are determined according to the AIC, allowing different lag lengths (from 0 to 8) on each variable. For the system of bivariate $(\Delta Y, x)$ equations to be used in the bootstrap, we adjust the coefficients of the OLS–estimated models for the small–sample bias that can plague time series models. Specifically, we use the bootstrap method proposed by Kilian (1998) to adjust the coefficients of the OLS–estimated models (based on 10,000 bootstrap draws) and then use the bias–adjusted forms as

the bootstrap DGP equations.

Bootstrapped time series on ΔY_t and x_t are generated by drawing with replacement from the sample residuals and using the autoregressive structures of the bias–adjusted models to iteratively construct data. The initial observations — observations preceding the sample of data used to estimate the models — necessitated by the lag structures of the estimated models, are selected by sampling from the actual data. In particular, following Stine (1987), among others, the initial observations are selected by picking one date at random and then taking the necessary number of initial observations in order from that date backward.[13]

In each of 999 bootstrap replications, the bootstrapped data are used to recursively estimate the restricted and unrestricted DMS forecasting models on which the sample results are based. The resulting forecasts are then used to calculate forecast test statistics. Critical values are simply computed as percentiles of the bootstrapped test statistics.

Overall, despite the parametric nature of this bootstrap procedure, its success in our results does not hinge on the bootstrap models being properly specified. For simplicity, the estimated models for ΔY and x are taken to be correctly specified in bootstrapping artificial data. However, those models may in fact be misspecified, because, as described above, their lag orders were determined with the sample data (each artificial data set). In this sense, our bootstrap is reflective of the various bootstrap approaches that have been used in studies such as Mark (1995), Kilian (1999), Rapach and Weber (2004), and Stock and Watson (2003). Therefore, if a simple, potentially–misspecified bootstrap proves reliable in our Monte Carlo experiments, it can be expected to be reliable in practice, in similar settings. All that said, it could be that nonparametric bootstrap approaches, such as moving block methods, would perform as well or better. But in light of Berkowitz and Kilian's (2000) conclusion that, for time series models, such methods are often dominated by parametric bootstraps, we leave nonparametric methods as a subject for future research on forecast evaluation.

[13]For example, suppose the model is a VAR(4) and the total sample consists of 144 observations, such that observations 1-4 serve as the initial observations and the regression sample is 5 through 144. Each artificial data set is constructed by: (i) picking a random date t_0 from a range of 5 through 144; (ii) setting the artificial observations 1-4 equal to the sample observations from dates $t_0 - 4$ through $t_0 - 1$; and (iii) constructing artificial observations 5-144 by using the VAR structure, resamples of the residuals (which span obs. 5–144), and the artificial initial observations 1-4.

4.3 Monte Carlo Results: Size

The results presented in Tables 1 and 2 indicate that, in some but not all cases, asymptotic critical values yield tests that are reasonably close to correctly sized. In particular, the MSE-F test compared against asymptotic critical values seems to have decent size in most settings. For example, as shown in Table 1's results for DGP-1, with $R = 100$ and $P = 40$, the size of the MSE-F test ranges from 10.0 percent for $\tau = 1$ to 12.4 percent for $\tau = 12$. Admittedly, though, in a few instances the size of the MSE-F test at longer horizons is subject to slightly larger distortions — such as size of 15.0 percent with DGP-1, $R = 60$, $P = 40$, and $\tau = 12$. Compared to MSE-F, the ENC-F test is subject to consistently larger size distortions when asymptotic critical values are used. For example, with DGP-1, $R = 100$ and $P = 40$, the size of the ENC-F test is roughly 15 percent for all horizons. The larger distortions in the ENC-F test are consistent with the one–step ahead results of Clark and McCracken (2001).

The performance of the MSE-T and ENC-T tests based on asymptotic critical values is generally mixed. The tests (MSE-T more so than ENC-T) can have decent size properties at short forecast horizons but are dramatically oversized at long horizons. For instance, as shown in Table 2's results for DGP-2 with $R = 100$ and $P = 40$, the MSE-T and ENC-T test sizes are 10.7 and 11.8 percent, respectively, for $\tau = 1$, but 23.9 and 29.7 percent for $\tau = 12$. At all but the shortest horizons, the sizes of the -T tests are usually greater than the sizes of the corresponding -F tests. The root of the problem in the longer-horizon performance of the MSE-T and ENC-T tests compared against asymptotic critical values seems to be imprecision in estimation of the HAC variance in the denominator of the test statistics.[14] In unreported simulations, the performance of these tests improved dramatically when R and P were increased significantly. Moreover, in applications in which the null forecasting model is a random walk, Clark and West (2004) find that using the HAC estimator of Hodrick (1992) rather than the common Newey and West (1987) estimator greatly improves the size of t-tests for equal MSE and forecast encompassing. Unfortunately, though, the estimator of Hodrick can only be applied when the null forecasting model takes a random

[14]Imprecision in the estimate of $S_{\tilde{h}\tilde{h}}$ used to construct the asymptotic critical values could be another source of difficulty. But we obtained results similar to those reported when we used an estimate of $S_{\tilde{h}\tilde{h}}$ based on a separate, very large sample of artificial data (rather than the small sample used in computing the test statistics themselves).

walk or "no change" form (and therefore has no estimated parameters).

Perhaps not surprisingly, using bootstrap critical values instead of asymptotic critical values yields better size results. Although the encompassing tests are sometimes modestly oversized at shorter horizons, the other tests are all consistently (reasonably) close to being correctly sized when based on bootstrap critical values. As shown in Table 1, for example, in the case of DGP-1, when $\tau = 4$, $R = 60$, and $P = 80$, the MSE-F, MSE-T, ENC-F, and ENC-T, statistics have size of 10.1, 10.5, 12.1, and 11.1 percent, respectively. Table 2 shows that, with DGP-2, $\tau = 4$, $R = 60$, and $P = 80$, the sizes of the tests (same order) are 10.4, 10.0, 11.5, and 11.4 percent, respectively. Using the bootstrap is particularly important for improving the small sample properties of the MSE-T and ENC-T tests at longer horizons — the bootstrapped critical values seem to reflect the imprecision in small sample estimates of the HAC variance that enters the test statistics.

In light of the past use of standard normal critical values in applied research applying t-tests for equal MSE to forecasts from nested models (recent examples include Clarida, et al. (2003) and Cheung, Chinn, and Pascual (2003)), a natural question is, how would using standard normal critical values affect inference under the null? As shown in Table 3 (we present results for just DGP-1 in the interest of brevity), in our experiment settings standard normal critical values can lead to serious under–rejection at short horizons and over–rejection at long horizons. For example, with $R = 100$ and $P = 40$, the MSE-T test (corresponding to the so–called Diebold–Mariano test) has size of about 4 percent for $\tau = 1$ and 2 but size of 16.1 percent for $\tau = 12$. The size of the ENC-T test is consistently higher, but shows the same pattern of rising sharply with the forecast horizon, such that the test is undersized or about correctly sized for shorter horizons but oversized for longer horizons.

In general, for standard normal critical values to provide reliable inference, the forecast horizon needs to be relatively short, and P/R needs to be quite small. Once the forecast horizon increases beyond a few periods, neither a standard normal approximation nor our asymptotic distribution yields reliable inference in finite samples; bootstrap methods are much more reliable. At short horizons, the standard normal approximation might be seen as acceptable for the ENC-T test, but not the MSE-T test. In the results reported in Table 3, the size of the ENC-T test for $\tau = 1$ and 2 ranges from 8.9 to 12.7 percent. The ENC-T statistic's performance is less favorable

for $\tau = 4$, with size rising to 15 percent. For horizons $\tau \leq 4$, the MSE-T test is undersized for all of the P/R settings reported, although less so for small P/R than large P/R. For longer horizons, the MSE-T test ranges from under- to over-sized, depending on the horizon and sample sizes.

At shorter horizons, how small does P/R need to be for the MSE-T test to be reliably compared against standard normal critical values? Even with $P/R = 40/200 = .2$, the MSE-T test has size of 5.3 percent for $\tau = 1$ and 6.6 percent for $\tau = 2$ (Table 3). Some additional simulations indicate that P/R needs to be less than .10 for standard normal critical values to be reliably used. For example, with DGP-1, $R = 400$, and $P = 40$ (such that $P/R = .1$) the empirical size of the MSE-T test compared against standard normal critical values is 6.2, 8.5, and 11.8 percent for $\tau = 1$, 2, and 4, respectively (using our asymptotic critical values yields a size of about 12 percent). Doubling R (so $P/R = .05$) makes the empirical size against standard normal critical values go up slightly, to 7.2, 9.8, and 13.2 percent for $\tau = 1$, 2, and 4, respectively. As noted in section 3.1, in many historical forecast evaluations, P/R is considerably larger than .10 or .05. As a result, the standard normal approximation seems unlikely to be accurate for the commonly–used MSE-T (or Diebold–Mariano) test.

4.4 Monte Carlo Results: Power

In evaluating power, we begin with results based on bootstrap critical values, because the bootstrap–based tests are, for the most part, about correctly sized. The bootstrap–based power results presented in Tables 4 and 5 indicate the test powers follow the same general ranking as in Clark and McCracken's (2001) Monte Carlo examination of tests based on one–step ahead forecasts: ENC-F > MSE-F, ENC-T > MSE-T. MSE-F is often more powerful than ENC-T, and sometimes much more so, but the ranking of these two tests varies with τ and the R, P setting. For example, Table 4 shows that, with DGP-1, $\tau = 4$, $R = 100$, and $P = 40$, the bootstrap–based powers of the MSE-F, MSE-T, ENC-F, and ENC-T tests are 53.1, 34.4, 70.2, and 47.8 percent, respectively. Clark and McCracken (2005) prove that, asymptotically, the MSE-F and ENC-F tests are more powerful than their t-type counterparts because, under the alternative hypothesis, the F–type tests diverge to infinity at a faster rate.

For both of the DGPs considered, power generally falls as τ rises, and the power differences among the tests tend to decline. With DGP-1, $R = 100$, and $P = 40$, for

example, Table 4 shows that the bootstrap–based power of the MSE-T test declines from 49.8 percent when $\tau = 1$ to 18.3 percent when $\tau = 12$. The power of the ENC-T declines more sharply as τ rises, from 73.8 percent when $\tau = 1$ to 21.3 percent when $\tau = 12$. As a result, the power advantage of ENC-T over MSE-T shrinks as the forecast horizon grows.[15] As might be expected, power tends to rise with both R and P. Given P, the powers of the tests tend to rise with R, more so for MSE-F and ENC-F than the other forecast tests. For example, Table 4 shows that, with DGP-1, $P = 40$, and $\tau = 2$, the bootstrap–based power of the MSE-F test increases from 63.8 percent when $R = 100$ to 72.9 percent when $R = 200$. Given R, increases in the number of forecast observations consistently lead to a rise in power. As reported in Table 5, with DGP-2, $R = 60$, and $\tau = 4$, the power of the MSE-F test rises from 49.7 percent when $P = 40$ to 68.5 percent when $P = 80$ and 79.0 percent when $P = 120$.

Power based on asymptotic critical values produces most of the same basic patterns, although the asymptotics–based powers of some of the tests can differ substantially from their bootstrap–based powers, reflecting the size distortions of the tests. For the one test that seems to have decent size properties across all forecast horizons when based on asymptotic critical values, the MSE-F test, power based on asymptotic critical values is quite close to the power estimates based on bootstrap methods. For example, with DGP-2, $R = 60$, and $P = 80$, the power of the MSE-F test based on asymptotic critical values ranges from 44.2 percent for $\tau = 12$ to 78.7 percent for $\tau = 1$, compared to the bootstrap–based powers of 40.5 to 79.2 percent (Table 5). For the other tests, often subject to size distortions at longer forecast horizons, power based on asymptotic critical values is generally greater than bootstrap–based power (the differences become especially large for the MSE-T and ENC-T tests at longer horizons, because, for these tests, size distortions rise sharply with the forecast horizon). In the same example, the power of the ENC-T test ranges from 42.1 to 93.6 percent with bootstrap critical values but 68.7 to 95.0 percent with asymptotic critical values.

[15]Based on prior experiments with other DGPs, it seems that the relationship of power to τ depends on the DGP in important ways, making generalizations difficult. Mark and Sul (2002) use local asymptotic analysis to show that certain DGP features will cause power to rise with the forecast horizon.

5 Application to Inflation Forecasting

In this section we use the tests and inference approaches described above to determine whether capacity utilization is useful for predicting core CPI inflation. Cecchetti (1995), Staiger, Stock, and Watson (1997), and Stock and Watson (1999, 2003) are recent examples of studies in the long literature on this basic question. Like these other studies, we examine out–of–sample forecasts to gauge the predictive content of capacity utilization.

Our quarterly data on the core CPI and capacity utilization in manufacturing span 1957:Q1 through 2004:Q3. After allowances for data differencing and a maximum of four data–determined lags, the sample period available for estimation of a 1–step ahead prediction model spans 1958:Q3–2004:Q3, for a total of 185 observations. We begin forecasting in 1976:Q1, so that $P = 115$.

Following the basic approach of Stock and Watson (1999, 2003), we treat inflation as having a unit root, and forecast a measure of the direct multi-step change in inflation as a function of lags of the change in quarterly inflation and lags of capacity utilization. In particular, using the notation of the last section, we make Y the log difference of the quarterly core CPI (scaled by 400 to make Y an annualized percentage change); ΔY is then the change in quarterly inflation. The predictand is $Y_{t+\tau}^{(\tau)} - Y_t$, where $Y_{t+\tau}^{(\tau)}$ denotes the average annual rate of price change from t to $t+\tau$. x denotes the rate of capacity utilization in manufacturing. The restricted model (model 1) is autoregressive — the multi-step change in inflation is a function of just lags of the one–period change in inflation. The unrestricted model (model 2) adds lags of capacity utilization to the set of regressors. In particular, the competing forecasting models take the forms of section 4.1.3's equations (12) and (13). For each forecast horizon, we use the AIC to determine the lag orders of (13), allowing different lag lengths for inflation and capacity utilization.[16] The baseline AR model (12) uses the inflation lag order selected for (13). The lag selection is based on just the in–sample portion of the data (1958–1975 model estimates).

We use both the asymptotic approach described in section 3.3 and the bootstrap approach described in section 4.2 to draw inferences on capacity utilization's predictive power for inflation. In bootstrapping, we use the full sample of data to estimate

[16]Results based on a fixed order of two lags of inflation and four lags of capacity utilization are similar.

vector autoregressive equations for the one–quarter change in inflation ΔY_t and capacity utilization x_t, imposing the null that utilization has no predictive power for inflation. The DGP equation for ΔY takes exactly the same form as the restricted forecasting model for $\tau = 1$ (but estimated with all available data). The lag orders of the capacity utilization equation are determined according to the AIC. The coefficients of the DGP equations are bias–adjusted with Kilian's (1998) procedure. Then bootstrapped time series on ΔY_t and x_t are generated by sampling the residuals and using the autoregressive structures of the bias–adjusted models to iteratively construct data.[17]

The results reported in Table 6 indicate that, over the 1976-2004 period, capacity utilization in manufacturing has significant predictive power for core inflation. As shown in the upper panel of the table, for all horizons considered, forecasts from the model with capacity utilization (Model 2) have a lower RMSE than forecasts from the autoregressive model (Model 1). The test statistics and p-values in the lower panel indicate capacity utilization's predictive content is statistically significant. Consistent with our Monte Carlo evidence that the F-type tests are more powerful than their t-type counterparts, the MSE-F and ENC-F p-values are generally lower than those of MSE-T and ENC-T. The tendency of the asymptotic p-values to be slightly lower than the bootstrap p-values is also consistent with the Monte Carlo evidence.

6 Conclusion

In this paper we first derive the limiting distributions of four tests of direct multi-step forecasts from linear regression models: the t–statistic for equal MSE developed by Diebold and Mariano (1995) and West (1996); the F–type test of equal MSE proposed by McCracken (2004); the t–statistic for encompassing developed in Harvey, Leybourne, and Newbold (1998) and West (2001); and the encompassing test proposed by Clark and McCracken (2001). We show that, when the number of observations used to generate initial estimates of the models and the number of forecast observations increase at the same rate, all of the tests have non-standard distribu-

[17] In light of the potential for conditional heteroskedasticity, in this application we slightly modify the bootstrap procedure used in the Monte Carlo analysis and use the wild bootstrap recommended by Goncalves and Kilian (2004). Instead of sampling from the residuals with replacement, we use artificial residuals that are the product of the sample residuals (kept in their original order) and an i.i.d. draw from the standard normal distribution.

tions. While these distributions can be free of nuisance parameters when the forecast horizon is one, they are not free of nuisance parameters for longer forecast horizons.

Using both our asymptotics and a simple model–based bootstrap for estimating appropriate critical values, we then conduct a range of Monte Carlo simulations to examine the finite–sample properties of the tests. These results indicate our asymptotic approximation yields good finite–sample size and power properties for some, but not all, of the tests considered. In general, the asymptotics seem to work well for McCracken's (2004) F–type test of equal MSE. A simple bootstrap works reasonably well for all tests. Finally, the encompassing test proposed by Clark and McCracken (2001) — the ENC-F statistic defined in equation (4) — is most powerful.

In the final part of our analysis, applying our tests shows that capacity utilization in manufacturing has significant predictive power for core inflation in the U.S. For out–of–sample forecasts over 1976-2004, all of the tests of equal forecast accuracy and encompassing indicate that capacity utilization improves forecasts of core inflation at all horizons.

7 Appendix: Proofs

The following notation will be used. For any $(m \times n)$ matrix G with elements $g_{i,j}$ and column vectors g_j let $vec(G)$ denote the $(mn \times 1)$ vector $\left[g_1^{'}, g_2^{'}, ..., g_n^{'}\right]^{'}$ and let $|G|$ denote $\max_{i,j}|g_{i,j}|$. For any sequence z_t, $\sum_t z_t$ denotes $\sum_{t=R}^{T-\tau} z_t$, S_{zz} denotes $\lim Var(P^{-1/2}\sum_t z_t)$ and $\sup_t |z_t|$ denotes $\sup_{R\leq t\leq T}|z_t|$.

For brevity, much of the extensive algebra involved in the proofs of Theorems 3.1–3.4 is relegated to a not-for-publication technical appendix, Clark and McCracken (2004). Before proceeding to the proofs we first provide an appendix Lemma.

Lemma A1: Under Assumptions 1, 2, and 4, $\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+\tau} \rightarrow_d \int_\lambda^1 \omega^{-1}W'(\omega)S_{\tilde{h}\tilde{h}}dW(\omega)$.

Proof of Lemma A1: The results are modifications of those in Hansen (1992). Using Hansen's notation, let the operator E_iX denote $E(X|\Im_i)$, where $\Im_t \equiv \sigma(T^{-1/2}\sum_{s=1}^i \tilde{h}_{2,s}, \tilde{h}_{2,i} : i \leq t, T \geq 1)$ is the smallest sigma-field containing the history of $\{T^{-1/2}\sum_{s=1}^t \tilde{h}_{2,s}, \tilde{h}_{2,t}\}\ \forall\ T$. Define $\varepsilon_{t+\tau} = \sum_{i=0}^\infty (E_i\tilde{h}_{2,t+\tau+i} - E_{i-1}\tilde{h}_{2,t+\tau+i})$ and $z_{t+\tau} = \sum_{i=1}^\infty E_i\tilde{h}_{2,t+\tau+i}$. Then $\tilde{h}_{2,t+\tau} = \varepsilon_{t+\tau} + z_{t+\tau-1} + z_{t+\tau}$.

In the above notation, $\sum_t (T/t)(T^{-1/2}\sum_{s=1}^{t-\tau}\tilde{h}_{2,s+\tau})'(T^{-1/2}\tilde{h}_{2,t+\tau})$

$$
\begin{aligned}
&= \sum_t (T/t)(T^{-1/2}\sum_{s=\tau+1}^{t+\tau-1}\tilde{h}_{2,s})'(T^{-1/2}\tilde{h}_{2,t+\tau}) - \sum_t (1/t)(\sum_{j=1}^{\tau-1}\tilde{h}_{2,t+j})'\tilde{h}_{2,t+\tau} \\
&= \sum_t (T/t)(T^{-1/2}\sum_{s=\tau+1}^{t+\tau-1}\tilde{h}_{2,s})'(T^{-1/2}\varepsilon_{t+\tau}) + \sum_t (1/t)(\sum_{s=\tau+1}^{t+\tau-1}\tilde{h}_{2,s})'(z_{t+\tau-1} - z_{t+\tau}) \\
&\quad - \sum_t (1/t)(\sum_{j=1}^{\tau-1}\tilde{h}_{2,t+j})'\tilde{h}_{2,t+\tau} \\
&= \sum_t (T/t)(T^{-1/2}\sum_{s=1}^{t-\tau}\tilde{h}_{2,s+\tau})'(T^{-1/2}\varepsilon_{t+\tau}) + R^{-1}(\sum_{s=1+\tau}^{R+\tau-2}\tilde{h}_{2,s}')z_{R+\tau-1} \\
&\quad - T^{-1}(\sum_{s=1+\tau}^{T+\tau-1}\tilde{h}_{2,s}')z_{T+\tau} - \sum_{t=R}^{T-1}(t^2+t)^{-1}(\sum_{s=1+\tau}^{t+\tau-1}\tilde{h}_{2,s})'z_{t+\tau} \\
&\quad + \sum_{t=R-1}^{T-1}(1/t)\tilde{h}_{2,t+\tau}'z_{t+\tau} - \sum_t (1/t)(\sum_{j=1}^{\tau-1}\tilde{h}_{2,t+j})'\tilde{h}_{2,t+\tau}.
\end{aligned}
$$

That $\sum_t (T/t)(T^{-1/2}\sum_{s=1}^{t-\tau}\tilde{h}_{2,s+\tau})'(T^{-1/2}\varepsilon_{t+\tau}) \rightarrow_d \int_\lambda^1 \omega^{-1}W'(\omega)S_{\tilde{h}\tilde{h}}dW(\omega)$ follows from Theorem 4.1 of Hansen (1992). Lemma A1 follows if the sum of the remaining terms is $o_p(1)$.

Consider the second and third right-hand side terms. Taking their absolute value we obtain

$|R^{-1}(\sum_{s=1+\tau}^{R+\tau-2}\tilde{h}_{2,s}')z_{R+\tau-1}| \leq (T/R)k_2|T^{-1/2}\sum_{s=1+\tau}^{R+\tau-2}\tilde{h}_{2,s}||T^{-1/2}z_{R+\tau-1}|$, and

$|R^{-1}(\sum_{s=1+\tau}^{T+\tau-1}\tilde{h}'_{2,s})z_{T+\tau}| \leq (T/R)k_2|T^{-1/2}\sum_{s=1+\tau}^{T+\tau-1}\tilde{h}_{2,s}||T^{-1/2}z_{T+\tau}|$.

Assumption 4 implies that T/R is bounded while Assumption 2 implies that both $|T^{-1/2}\sum_{s=1+\tau}^{R+\tau-2}\tilde{h}_{2,s}|$ and $|T^{-1/2}\sum_{s=1+\tau}^{T+\tau-1}\tilde{h}_{2,s}|$ are $O_p(1)$. That the second and third right-hand side terms are $o_p(1)$ follows from (A.3) of Hansen (1992) wherein he shows that both $|T^{-1/2}z_{R+\tau-1}|$ and $|T^{-1/2}z_{T+t}|$ are $o_p(1)$.

Consider the fourth right-hand side term. Taking its absolute value we obtain

$$|\sum_{t=R}^{T-1}(t^2+t)^{-1}(\sum_{s=1+\tau}^{t+\tau-1}\tilde{h}_{2,s})'z_{t+\tau}|$$

$$\leq [(T-1-R)/(R^2+R)]k_2(\sup_t|T^{-1/2}\sum_{s=1+\tau}^{t+\tau-1}\tilde{h}_{2,s}|)(\sup_{t\leq T}|T^{-1/2}z_{t+\tau}|).$$

Assumption 2 implies $(\sup_t|T^{-1/2}\sum_{s=1+\tau}^{t+\tau-1}\tilde{h}_{2,s}|) = O_p(1)$. That $(\sup_{t\leq T}|T^{-1/2}z_{t+\tau}|) = o_p(1)$ follows from (A.3) of Hansen (1992). The result follows since by Assumption 4, $(T-1-R)/(R^2+R) = o_p(1)$.

Consider the fifth right-hand side term. We show that it converges in probability to $-\ln(\lambda)\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau}$. Rearranging terms we obtain
$\sum_{t=R-1}^{T-1}(1/t)(\tilde{h}'_{2,t+\tau}z_{t+\tau} - E\tilde{h}'_{2,t+\tau}z_{t+\tau}) + \sum_{t=R-1}^{T-1}(1/t)(E\tilde{h}'_{2,t+\tau}z_{t+\tau})$. Since
$E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j} = 0$ for all $j \geq \tau$ it is clear that $\sum_{t=R-1}^{T-1}(1/t)(E\tilde{h}'_{2,t+\tau}z_{t+\tau})$
$= \sum_{t=R-1}^{T-1}(1/t)(E\tilde{h}'_{2,t+\tau}[\sum_{i=1}^{\infty}E_i\tilde{h}_{2,t+\tau+i}]) = (T^{-1}\sum_{t=R-1}^{T-1}(T/t))(\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau})$.
Since for large enough T, $T^{-1}\sum_{t=R}^{T}(T/t) \sim \int_{\lambda}^{1}\omega^{-1}d\omega = -\ln(\lambda)$, the result will follow if $\sum_{t=R-1}^{T-1}(1/t)(\tilde{h}'_{2,t+\tau}z_{t+\tau} - E\tilde{h}'_{2,t+\tau}z_{t+\tau}) = T^{-1}\sum_{t=R-1}^{T-1}(T/t)(\tilde{h}'_{2,t+\tau}z_{t+\tau} - E\tilde{h}'_{2,t+\tau}z_{t+\tau})$
$= o_p(1)$. If we define $U_{Tt} \equiv T/t$ and $e_t \equiv \tilde{h}'_{2,t+\tau}z_{t+\tau} - E\tilde{h}'_{2,t+\tau}z_{t+\tau}$ then the result follows from Theorem 3.2 of Hansen (1992).

Because of the minus sign, the proof will be complete if the final right-hand side term converges in probability to $-\ln(\lambda)\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau}$. Rearranging terms yields $\sum_t(1/t)(\sum_{j=1}^{\tau-1}\tilde{h}_{2,t+j})'\tilde{h}_{2,t+\tau} = T^{-1/2}\sum_{j=1}^{\tau-1}T^{-1/2}\sum_t(T/t)(\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau} - E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau}) +$
$(T^{-1}\sum_t(T/t))(\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau})$. Given Assumption 2, Corollary 29.11 of Davidson (1994) implies that $\sum_{j=1}^{\tau-1}T^{-1/2}\sum_t(T/t)(\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau} - E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau}) = O_p(1)$. Since $T^{-1/2} = o_p(1)$ the result is obtained because $(T^{-1}\sum_t(T/t))(\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau}) =$
$-\ln(\lambda)\sum_{j=1}^{\tau-1}E\tilde{h}'_{2,t+j}\tilde{h}_{2,t+\tau} + o_p(1)$ was established in the preceding paragraph.

Proof of Theorem 3.1: (a) Given Theorem 3.4 and the Continuous Mapping Theorem it suffices to show that $P\sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\hat{\Gamma}_{cc}(j) \rightarrow_d \sigma^4\Gamma_3$. Lengthy algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply that $P\hat{\Gamma}_{cc}(j) = \sigma^4\sum_t\tilde{H}'_2(t)[E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j}]\tilde{H}_2(t) +$

$o_p(1)$. Substitution, and the fact that $\bar{j}$ is finite provides $P\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{cc}(j) =$

$$\begin{aligned}
& \sigma^4 \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)[\sum_t \tilde{H}_2'(t)[E\tilde{h}_{2,t+\tau}\tilde{h}_{2,t+\tau-j}']\tilde{H}_2(t)] + o_p(1) \\
= \; & \sigma^4 \sum_t \tilde{H}_2'(t)[\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}_{2,t+\tau-j}')]\tilde{H}_2(t) + o_p(1) \\
= \; & \sigma^4(T^{-1}\sum_t [T^{1/2}\tilde{H}_2'(t) \otimes T^{1/2}\tilde{H}_2'(t)])vec[\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}_{2,t+\tau-j}')] + o_p(1).
\end{aligned}$$

Given Assumption 3, $\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}_{2,t+\tau-j}') \rightarrow S_{\tilde{h}\tilde{h}}$. Since Assumption 2 and Corollary 29.19 of Davidson (1994) suffice for $T^{1/2}\tilde{H}_2(t) \Rightarrow \omega^{-1}S_{\tilde{h}\tilde{h}}^{1/2}W(\omega)$, the Continuous Mapping Theorem implies $T^{-1}\sum_t T^{1/2}\tilde{H}_2'(t) \otimes T^{1/2}\tilde{H}_2'(t) \rightarrow_d$ $\int_\lambda^1 \omega^{-2}[W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2} \otimes W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2}]d\omega$. Since $(\int_\lambda^1 \omega^{-2}[W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2} \otimes W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2}]d\omega)vec[S_{\tilde{h}\tilde{h}}] =$ Γ_3, we obtain the desired result.

(b) First consider the numerator of ENC-T. Additional algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply

$$\sum_t (\hat{u}_{1,t+\tau}^2 - \hat{u}_{1,t+\tau}\hat{u}_{2,t+\tau}) = (P/R)^{1/2}\sigma^2[R^{1/2}\tilde{H}_2'(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}] + o_p((P/R)^{1/2}).$$

Now consider the denominator of ENC-T. Similar algebra implies that

$$\begin{aligned}
P\hat{\Gamma}_{cc}(j) &= \sum_{t=R+j}^{T-\tau} (\hat{u}_{1,t+\tau}^2 - \hat{u}_{1,t+\tau}\hat{u}_{2,t+\tau} - \bar{c})(\hat{u}_{1,t+\tau-j}^2 - \hat{u}_{1,t+\tau-j}\hat{u}_{2,t+\tau-j} - \bar{c}) \\
&= (P/R)\sigma^4[R^{1/2}\tilde{H}_2'(R)][E\tilde{h}_{2,t+\tau}\tilde{h}_{2,t+\tau-j}'][R^{1/2}\tilde{H}_2(R)] + o_p(P/R).
\end{aligned}$$

Substitution, and using the fact that $\bar{j}$ is finite then provides

$$\begin{aligned}
\text{ENC-T} &= \frac{(P/R)^{1/2}\sigma^2[R^{1/2}\tilde{H}_2'(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}] + o_p((P/R)^{1/2})}{[(P/R)\sigma^4[R^{1/2}\tilde{H}_2'(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}_{t+\tau-j}')][R^{1/2}\tilde{H}_2(R)] + o_p(P/R)]^{1/2}} \\
&= \frac{[R^{1/2}\tilde{H}_2'(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}] + o_p(1)}{[[R^{1/2}\tilde{H}_2'(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}_{t+\tau-j}')][R^{1/2}\tilde{H}_2(R)] + o_p(1)]^{1/2}} \\
&= \frac{[R^{1/2}\tilde{H}_2'(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}]}{[[R^{1/2}\tilde{H}_2'(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}_{t+\tau-j}')][R^{1/2}\tilde{H}_2(R)]]^{1/2}} + o_p(1).
\end{aligned}$$

Given Assumption 2, Corollary 29.19 of Davidson (1994) suffices for $(P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}', R^{1/2}\tilde{H}_2'(R))' \rightarrow_d (V_1'S_{\tilde{h}\tilde{h}}^{1/2}, V_0'S_{\tilde{h}\tilde{h}}^{1/2})\prime$ for independent $(k \times 1)$ standard normal vectors V_0 and V_1. Given Assumption 3, we know that

$\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j}) \rightarrow S_{\tilde{h}\tilde{h}}$. The result follows immediately from the Continuous Mapping Theorem.

Proof of Theorem 3.2: (a) Given Theorem 3.3 and the Continuous Mapping Theorem it suffices to show that $\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{dd}(j) \rightarrow_d 4\sigma^4\Gamma_3$. Extensive algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply that $P\hat{\Gamma}_{dd}(j) = 4\sigma^4 \sum_t \tilde{H}'_2(t)[E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j}]\tilde{H}_2(t) + o_p(1)$. Substitution, and the fact that $\bar{j}$ is finite provides $P\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{dd}(j) =$

$$
\begin{aligned}
& 4\sigma^4 \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)[\sum_t \tilde{H}'_2(t)[E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j}]\tilde{H}_2(t)] + o_p(1) \\
= \; & 4\sigma^4 \sum_t \tilde{H}'_2(t)[\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j})]\tilde{H}_2(t) + o_p(1) \\
= \; & 4\sigma^4(T^{-1}\sum_t [T^{1/2}\tilde{H}'_2(t) \otimes T^{1/2}\tilde{H}'_2(t)])vec[\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j})] + o_p(1).
\end{aligned}
$$

The result follows immediately from the proof of Theorem 3.1 (a).

(b) First consider the numerator of MSE-T. Extensive algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply

$$
\sum_t (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) = 2(P/R)^{1/2}\sigma^2[R^{1/2}\tilde{H}'_2(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+\tau}] + o_p((P/R)^{1/2}).
$$

Now consider the denominator of MSE-T. Similar algebra implies that

$$
\begin{aligned}
P\hat{\Gamma}_{dd}(j) &= \sum_{t=R+j}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau} - \bar{d})(\hat{u}^2_{1,t+\tau-j} - \hat{u}^2_{2,t+\tau-j} - \bar{d}) \\
&= 4(P/R)\sigma^4[R^{1/2}\tilde{H}'_2(R)][E\tilde{h}_{2,t+\tau}\tilde{h}'_{2,t+\tau-j}][R^{1/2}\tilde{H}_2(R)] + o_p(P/R).
\end{aligned}
$$

Substitution, and using the fact that $\bar{j}$ is finite then provides

$$
\begin{aligned}
\text{MSE-T} &= \frac{2(P/R)^{1/2}\sigma^2[R^{1/2}\tilde{H}'_2(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+1}] + o_p((P/R)^{1/2})}{[4(P/R)\sigma^4[R^{1/2}\tilde{H}'_2(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-j})][R^{1/2}\tilde{H}_2(R)] + o_p(P/R)]^{1/2}} \\
&= \frac{[R^{1/2}\tilde{H}'_2(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+1}] + o_p(1)}{[[R^{1/2}\tilde{H}'_2(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-j})][R^{1/2}\tilde{H}_2(R)] + o_p(1)]^{1/2}} \\
&= \frac{[R^{1/2}\tilde{H}'_2(R)][P^{-1/2}\sum_t \tilde{h}_{2,t+1}]}{[[R^{1/2}\tilde{H}'_2(R)][\sum_{j=-\bar{j}}^{\bar{j}} K(j/M)(E\tilde{h}_{t+\tau}\tilde{h}'_{t+\tau-j})][R^{1/2}\tilde{H}_2(R)]]^{1/2}} + o_p(1).
\end{aligned}
$$

The result follows immediately from the proof of Theorem 3.1 (b).

Proof of Theorem 3.3: (a) That $P^{-1}\sum_t \hat{u}^2_{2,t+\tau} \rightarrow_p \sigma^2$ follows from Theorem 4.1 of West (1996). Extensive algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply that

$\sum_t (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) = 2\sigma^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+\tau} - \sigma^2 T^{-1} \sum_t (T^{1/2}\tilde{H}_2'(t))(T^{1/2}\tilde{H}_2'(t)) + o_p(1)$. Since Assumption 2 and Corollary 29.19 of Davidson (1994) suffice for $T^{1/2}\tilde{H}_2(t) \Rightarrow \omega^{-1} S_{\tilde{h}\tilde{h}}^{1/2} W(\omega)$, the Continuous Mapping Theorem implies $T^{-1} \sum_t (T^{1/2}\tilde{H}_2'(t))(T^{1/2}\tilde{H}_2'(t)) \rightarrow_d \Gamma_2$. The result then follows from Lemma A1.

(b) That $P^{-1} \sum_t \hat{u}_{2,t+\tau}^2 \rightarrow_p \sigma^2$ follows from Theorem 4.1 of West (1996). Detailed algebra (see Clark and McCracken (2004)) and the definition of $\tilde{h}_{2,t+\tau}$ imply that $\sum_t (\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) = 2\sigma^2 (P/R)^{1/2} [R^{1/2}\tilde{H}_2'(R)][P^{-1/2} \sum_t \tilde{h}_{2,t+\tau}] + o_p((P/R)^{1/2})$. Given Assumption 2, Corollary 29.19 of Davidson (1994) suffices for $(P^{-1/2} \sum_t \tilde{h}_{2,t+\tau}', R^{1/2}\tilde{H}_2'(R))' \rightarrow_d (V_1' S_{\tilde{h}\tilde{h}}^{1/2}, V_0' S_{\tilde{h}\tilde{h}}^{1/2})'$ for the independent $(k_2 \times 1)$ standard normal vectors V_0 and V_1 from Theorem 3.1. Scaling by $(R/P)^{1/2}$ provides the desired result.

Proof of Theorem 3.4: (a) That $P^{-1} \sum_t \hat{u}_{2,t+\tau}^2 \rightarrow_p \sigma^2$ follows from Theorem 4.1 of West (1996). Lengthy algebra and the definition of $\tilde{h}_{2,t+\tau}$ imply that $\sum_t (\hat{u}_{1,t+\tau}^2 - \hat{u}_{1,t+\tau}\hat{u}_{2,t+\tau}) = \sigma^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+\tau} + o_p(1)$. The result follows from Lemma A1.

(b) That $P^{-1} \sum_t \hat{u}_{2,t+\tau}^2 \rightarrow_p \sigma^2$ follows from Theorem 4.1 of West (1996). Extensive algebra (see Clark and McCracken (2004)) and the definition of $\tilde{h}_{2,t+\tau}$ imply that $\sum_t (\hat{u}_{1,t+\tau}^2 - \hat{u}_{1,t+\tau}\hat{u}_{2,t+\tau}) = \sigma^2 (P/R)^{1/2} [R^{1/2}\tilde{H}_2'(R)][P^{-1/2} \sum_t \tilde{h}_{2,t+\tau}] + o_p((P/R)^{1/2})$. The result follows immediately from the proof of Theorem 3.3 (b).

References

Berkowitz, Jeremy, and Lutz Kilian, 2000, "Recent Developments in Bootstrapping Time Series," *Econometric Reviews* 19, pp. 1-48.

Cecchetti, Stephen G., 1995, "Inflation Indicators and Inflation Policy," *NBER Macroeconomics Annual*, pp. 189-219.

Chao, John, Valentina Corradi, and Norman R. Swanson, 2001, "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics* 5, pp. 598-620.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual, 2003, "Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive?" *Journal of International Money and Finance*, forthcoming.

Chevillon, Guillaume, and David F. Hendry, 2004, "Non-Parametric Direct Multi-Step Estimation for Forecasting Economic Processes," *International Journal of Forecasting*, forthcoming.

Clarida, Richard H., Lucio Sarno, Mark P. Taylor, and Giorgio Valente, 2003, "The Out–of–Sample Success of Term Structure Models as Exchange Rate Predictors: A Step Beyond," *Journal of International Economics* 60, pp. 61-83.

Clark, Todd E., 2004, "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting* 23, pp. 115-39.

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

Clark, Todd E., and Michael W. McCracken, 2004, "Technical Appendix to 'Evaluating Long–Horizon Forecasts'," manuscript, available at *www.kansascityfed.org/Econres/staff/tec.htm.*

Clark, Todd E., and Michael W. McCracken, 2005, "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics* 124, pp. 1-31. ' Clark, Todd E., and Kenneth D. West, 2004, "Using Out–of–Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics*, forthcoming.

Clements, Michael P., and David F. Hendry, "Multi–Step Estimation for Forecasting," *Oxford Bulletin of Economics and Statistics* 58, pp. 657-84.

Corradi, Valentina, and Norman R. Swanson, 2002, "A Consistent Test for Nonlinear Out–of–Sample Predictive Accuracy," *Journal of Econometrics* 110, pp. 353-81.

Corradi, Valentina, Norman R. Swanson, and Claudia Olivetti, 2001, "Predictive

Ability with Cointegrated Variables," *Journal of Econometrics* 105, pp. 315-58.

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

Diebold, Francis X., and Canlin Li, 2004, "Forecasting the Term Structure of Government Bond Yields," *Journal of Econometrics*, forthcoming.

Diebold, Francis X., and Roberto S. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, pp. 253-63.

Ericsson, Neil R., 1992, "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration," *Journal of Policy Modeling* 14, pp. 465-95.

Estrella, Arturo, and Gikas A. Hardouvelis, 1991, "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance* 46, pp. 555-76.

Estrella, Arturo, Anthony P. Rodrigues, and Sebastian Schich, 2003, "How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States," *Review of Economics and Statistics* 85, pp. 629-44.

Gilbert, Scott, 2001, "Sampling Schemes and Tests of Regression Models," manuscript, Southern Illinois University-Carbondale.

Goncalves, Silvia, and Lutz Kilian, 2004, "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics* 123, pp. 89-120.

Granger, C.W.J., and Yongil Jeon, 2004, "Forecasting Performance of Information Criteria with Many Macro Series," *Journal of Applied Statistics* 31, pp. 1227-40.

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

Groen, Jan J.J., 1999, "Long Horizon Predictability of Exchange Rates: Is It for Real?" *Empirical Economics* 24, pp. 451-469.

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

Harvey, David I., Stephen J. Leybourne, and Paul Newbold, 1998, "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16, pp. 254-59.

Hodrick, Robert J., 1992, "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," *Review of Financial Studies* 5, pp. 357-86.

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

Kilian, Lutz, 1998, "Small–Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics* 80, pp. 218-30.

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

Kilian, Lutz, and Mark P. Taylor, 2003, "Why Is It So Difficult to Beat the Random Walk Forecast of Exchange Rates?" *Journal of International Economics* 60, pp. 85-107.

Marcellino, Massimiliano, 2002, "Instability and Non–Linearity in the EMU," IGIER working paper no. 211.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson, 2004, "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," manuscript, Princeton University.

Mark, Nelson C., 1995, "Exchange Rates and Fundamentals: Evidence on Long–Horizon Predictability," *American Economic Review* 85, pp. 201-18.

Mark, Nelson C., and Donggyu Sul, 2002, "Asymptotic Power Advantages of Long–Horizon Regressions," manuscript, Ohio State University.

McCracken, Michael W., 2004, "Asymptotics for Out–of–Sample Tests of Causality," manuscript, University of Missouri.

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

Newey, Whitney K., and Kenneth D. West, 1987, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, pp. 703-08.

Orphanides, Athanasios, and Simon van Norden, 2004, "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," *Journal of Money, Credit, and Banking*, forthcoming.

Qi, Min, and Jangru Wu, 2003, "Nonlinear Prediction of Exchange Rates with Monetary Fundamentals," *Journal of Empirical Finance* 10, pp. 623-40.

Rapach, David E., and Christian E. Weber, 2004, "Financial Variables and the Sim-

ulated Out–of–Sample Forecastability of U.S. Output Growth Since 1985: An Encompassing Approach," *Economic Inquiry* 42, pp. 717-38.

Rossi, Barbara, 2001, "Optimal Tests for Nested Model Selection with Underlying Parameter Instability," manuscript, Duke University.

Schorfheide, Frank, 2003, "VAR Forecasting Under Misspecification," *Journal of Econometrics*, forthcoming.

Shintani, Mototsugu, 2004, "Nonlinear Forecasting Analysis Using Diffusion Indexes: An Application to Japan," *Journal of Money, Credit, and Banking*, forthcoming.

Staiger, Douglas, James H. Stock and Mark W. Watson, 1997, "The NAIRU, Unemployment and Monetary Policy, *Journal of Economic Perspectives* 11, pp. 33-49.

Stine, Robert A., 1987, "Estimating Properties of Autoregressive Forecasts," *Journal of the American Statistical Association* 82, pp. 1072-78.

Stock, James H., and Mark W. Watson, 1999, "Forecasting Inflation," *Journal of Monetary Economics* 44, pp. 293-335.

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

Vuong, Quang H., 1989, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica* 57, pp. 307-33.

West, Kenneth D., 1996, "Asymptotic Inference About Predictive Ability," *Econometrica* 64, pp. 1067-84.

West, Kenneth D., 2001, "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics* 19, pp. 29-33.

West, Kenneth D., 2005, "Forecast Evaluation," in *Handbook of Economic Forecasting*, Elliott, Graham, Granger, Clive W.J., and Timmermann, Allan, eds., forthcoming.

West, Kenneth D., and Michael W. McCracken, 1998, "Regression–Based Tests of Predictive Ability," *International Economic Review* 39, pp. 817-40.

Table 1: Monte Carlo Results on Size, DGP-1

	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$
	$R=60$, $P=40$					$R=100$, $P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.114	.119	.125	.132	.150	.100	.103	.113	.125	.124
MSE-T	.116	.130	.142	.183	.239	.112	.123	.154	.205	.230
ENC-F	.170	.172	.170	.176	.180	.147	.146	.152	.143	.144
ENC-T	.141	.181	.205	.286	.350	.129	.159	.193	.261	.297
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.108	.094	.093	.090	.121	.094	.088	.091	.090	.092
MSE-T	.102	.094	.080	.084	.099	.090	.090	.080	.090	.083
ENC-F	.142	.126	.109	.116	.136	.138	.112	.113	.093	.103
ENC-T	.119	.102	.096	.092	.105	.101	.096	.085	.103	.091
	$R=60$, $P=80$					$R=100$, $P=80$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.132	.095	.114	.117	.123	.120	.119	.135	.136	.126
MSE-T	.125	.107	.122	.137	.146	.119	.120	.142	.163	.157
ENC-F	.181	.138	.155	.161	.140	.169	.169	.166	.161	.153
ENC-T	.177	.162	.183	.195	.217	.145	.163	.186	.215	.226
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.134	.098	.101	.091	.096	.121	.111	.110	.098	.098
MSE-T	.126	.096	.105	.087	.098	.109	.098	.095	.086	.086
ENC-F	.161	.124	.121	.103	.099	.156	.133	.129	.122	.108
ENC-T	.159	.109	.111	.096	.096	.132	.116	.112	.103	.089
	$R=60$, $P=120$					$R=200$, $P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.107	.094	.087	.104	.114	.113	.103	.114	.132	.139
MSE-T	.105	.087	.084	.108	.120	.114	.138	.167	.231	.270
ENC-F	.182	.157	.161	.161	.149	.139	.129	.127	.140	.136
ENC-T	.163	.146	.151	.188	.207	.131	.154	.185	.243	.302
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.111	.089	.083	.092	.100	.109	.089	.088	.094	.104
MSE-T	.114	.085	.076	.097	.090	.092	.083	.087	.085	.086
ENC-F	.167	.124	.115	.117	.104	.123	.107	.099	.098	.103
ENC-T	.148	.111	.099	.115	.099	.099	.091	.088	.086	.087

Notes:
1. The data generating process is defined in equation (6).
2. For each artificial data set, forecasts of $Y_{t+\tau} - Y_t$ are formed recursively using estimates of equations (10) and (11). These forecasts are then used to form the indicated test statistics, defined in Section 3. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics are compared against asymptotic and bootstrapped critical values, using a significance level of 10%. Sections 3.3 and 4.2 describe the asymptotic and bootstrap procedures.
4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 999.

Table 2: Monte Carlo Results on Size, DGP-2

	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$
	$R=60$, $P=40$					$R=100$, $P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.094	.107	.131	.126	.118	.098	.111	.124	.122	.124
MSE-T	.096	.119	.144	.183	.203	.107	.122	.153	.192	.239
ENC-F	.137	.159	.174	.167	.146	.136	.149	.164	.153	.144
ENC-T	.125	.162	.213	.268	.302	.118	.167	.205	.250	.297
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.095	.094	.101	.092	.089	.096	.095	.104	.092	.090
MSE-T	.091	.100	.114	.100	.097	.088	.082	.093	.088	.086
ENC-F	.109	.101	.116	.109	.091	.125	.125	.114	.113	.106
ENC-T	.108	.106	.116	.100	.096	.095	.105	.111	.097	.098
	$R=60$, $P=80$					$R=100$, $P=80$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.077	.078	.114	.101	.111	.095	.107	.114	.115	.131
MSE-T	.074	.085	.111	.116	.138	.087	.100	.112	.134	.164
ENC-F	.119	.134	.160	.137	.138	.131	.149	.158	.159	.150
ENC-T	.109	.137	.179	.198	.224	.117	.147	.168	.209	.233
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.079	.084	.104	.085	.091	.091	.101	.094	.090	.101
MSE-T	.083	.090	.100	.093	.088	.084	.083	.084	.081	.097
ENC-F	.102	.098	.115	.089	.091	.118	.124	.121	.112	.107
ENC-T	.087	.100	.114	.087	.089	.098	.114	.108	.111	.107
	$R=60$, $P=120$					$R=200$, $P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.062	.069	.091	.095	.110	.103	.112	.125	.118	.115
MSE-T	.067	.070	.095	.101	.119	.115	.139	.169	.224	.270
ENC-F	.115	.144	.149	.160	.155	.128	.141	.133	.133	.125
ENC-T	.100	.129	.146	.172	.196	.125	.161	.194	.257	.314
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.069	.072	.087	.091	.103	.101	.098	.096	.094	.088
MSE-T	.078	.076	.094	.094	.106	.092	.088	.092	.097	.105
ENC-F	.103	.104	.107	.095	.098	.115	.114	.097	.095	.090
ENC-T	.092	.100	.100	.090	.092	.099	.098	.094	.097	.099

Notes:

1. The data generating process is defined in equation (7).

2. For each artificial data set, forecasts of $Y_{t+\tau}^{(\tau)} - Y_t$ are formed recursively using estimates of equations (12) and (13). These forecasts are then used to form the indicated test statistics, defined in Section 3. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.

3. In each Monte Carlo replication, the simulated test statistics are compared against asymptotic and bootstrapped critical values, using a significance level of 10%. Sections 3.3 and 4.2 describe the asymptotic and bootstrap procedures.

4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 999.

Table 3: Monte Carlo Results on the Size of Tests Based on Standard Normal Critical Values, DGP-1

	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 8$	$\tau = 12$	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 8$	$\tau = 12$
	$R = 60,\ P = 40$					$R = 100,\ P = 40$				
MSE-T	.022	.035	.057	.089	.163	.040	.045	.069	.113	.161
ENC-T	.107	.127	.154	.226	.299	.093	.121	.153	.206	.259
	$R = 60,\ P = 80$					$R = 100,\ P = 80$				
MSE-T	.025	.020	.031	.040	.063	.029	.034	.047	.051	.076
ENC-T	.123	.111	.136	.157	.174	.104	.125	.130	.155	.165
	$R = 60,\ P = 120$					$R = 200,\ P = 40$				
MSE-T	.014	.021	.018	.028	.034	.053	.066	.090	.144	.199
ENC-T	.115	.105	.107	.142	.155	.089	.120	.139	.209	.259

Notes:
1. The data generating process is defined in equation (6).
2. For each artificial data set, forecasts of $Y_{t+\tau} - Y_t$ are formed recursively using estimates of equations (10) and (11). These forecasts are then used to form the indicated test statistics, defined in Section 3. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics (the same as those used in the results in Table 1) are compared against standard normal critical values (10%).
4. The number of Monte Carlo simulations is 1000.

Table 4: Monte Carlo Results on Power, DGP-1

	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$
	$R=60,\ P=40$					$R=100,\ P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.610	.551	.492	.399	.316	.652	.632	.548	.444	.343
MSE-T	.537	.525	.481	.443	.415	.531	.554	.524	.490	.457
ENC-F	.841	.776	.647	.475	.344	.903	.870	.718	.555	.407
ENC-T	.754	.735	.693	.622	.566	.782	.787	.728	.638	.598
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.602	.574	.466	.361	.247	.642	.638	.531	.429	.290
MSE-T	.520	.451	.353	.250	.193	.498	.465	.344	.248	.183
ENC-F	.807	.749	.593	.391	.263	.886	.857	.702	.514	.339
ENC-T	.717	.610	.456	.283	.212	.738	.666	.478	.307	.213
	$R=60,\ P=80$					$R=100,\ P=80$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.751	.721	.619	.522	.400	.817	.773	.689	.573	.453
MSE-T	.733	.688	.599	.536	.437	.752	.719	.640	.556	.474
ENC-F	.950	.914	.816	.650	.447	.973	.958	.883	.724	.564
ENC-T	.924	.901	.819	.721	.581	.947	.938	.870	.757	.661
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.755	.739	.613	.509	.367	.821	.775	.684	.546	.417
MSE-T	.740	.676	.542	.435	.314	.751	.675	.548	.431	.318
ENC-F	.938	.909	.792	.589	.372	.970	.959	.863	.688	.510
ENC-T	.904	.851	.707	.522	.353	.941	.889	.750	.550	.386
	$R=60,\ P=120$					$R=200,\ P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.844	.810	.723	.587	.467	.744	.729	.634	.526	.398
MSE-T	.840	.793	.703	.573	.483	.542	.589	.539	.518	.479
ENC-F	.974	.956	.895	.713	.536	.946	.929	.811	.656	.498
ENC-T	.962	.956	.891	.749	.629	.827	.848	.738	.684	.629
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.848	.815	.719	.580	.446	.738	.729	.633	.520	.379
MSE-T	.851	.804	.669	.542	.408	.511	.481	.352	.254	.174
ENC-F	.969	.953	.865	.641	.471	.938	.927	.809	.635	.452
ENC-T	.962	.932	.811	.607	.438	.791	.728	.515	.330	.229

Notes:
1. The data generating process is defined in equations (6) and (8).
2. For each artificial data set, forecasts of $Y_{t+\tau} - Y_t$ are formed recursively using estimates of equations (10) and (11). These forecasts are then used to form the indicated test statistics, defined in Section 3. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics are compared against asymptotic and bootstrapped critical values, using a significance level of 10%. Sections 3.3 and 4.2 describe the asymptotic and bootstrap procedures.
4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 999.

Table 5: Monte Carlo Results on Power, DGP-2

	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$	$\tau=1$	$\tau=2$	$\tau=4$	$\tau=8$	$\tau=12$
	$R=60,\ P=40$					$R=100,\ P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.617	.520	.507	.368	.283	.708	.611	.586	.463	.367
MSE-T	.546	.481	.502	.422	.377	.585	.544	.558	.510	.478
ENC-F	.856	.755	.700	.481	.326	.923	.834	.784	.622	.466
ENC-T	.786	.730	.729	.628	.559	.838	.766	.783	.722	.667
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.624	.523	.497	.345	.239	.704	.612	.581	.455	.325
MSE-T	.545	.437	.413	.273	.195	.563	.461	.411	.314	.231
ENC-F	.838	.702	.626	.379	.233	.914	.815	.747	.562	.387
ENC-T	.755	.587	.523	.316	.214	.806	.664	.575	.397	.274
	$R=60,\ P=80$					$R=100,\ P=80$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.787	.707	.688	.567	.442	.849	.775	.753	.625	.518
MSE-T	.762	.682	.667	.568	.480	.802	.723	.715	.612	.545
ENC-F	.973	.921	.873	.738	.535	.996	.959	.926	.813	.644
ENC-T	.950	.900	.880	.787	.687	.978	.940	.907	.843	.762
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.792	.722	.685	.554	.405	.847	.784	.747	.610	.480
MSE-T	.779	.692	.649	.509	.381	.802	.704	.662	.513	.398
ENC-F	.960	.904	.838	.654	.424	.995	.950	.908	.769	.565
ENC-T	.936	.857	.797	.618	.421	.967	.896	.846	.668	.516
	$R=60,\ P=120$					$R=200,\ P=40$				
	Using asymptotic critical values					*Using asymptotic critical values*				
MSE-F	.896	.809	.780	.667	.548	.797	.739	.700	.565	.458
MSE-T	.896	.795	.770	.658	.553	.627	.595	.604	.565	.533
ENC-F	.987	.969	.948	.817	.641	.971	.922	.889	.744	.589
ENC-T	.982	.962	.943	.866	.750	.898	.851	.843	.769	.732
	Using bootstrapped critical values					*Using bootstrapped critical values*				
MSE-F	.903	.823	.790	.672	.534	.790	.737	.703	.569	.441
MSE-T	.907	.820	.784	.645	.520	.598	.499	.445	.284	.216
ENC-F	.986	.960	.928	.745	.542	.965	.927	.886	.725	.559
ENC-T	.981	.944	.907	.766	.556	.878	.759	.664	.426	.309

Notes:

1. The data generating process is defined in equations (7) and (9).

2. For each artificial data set, forecasts of $Y_{t+\tau}^{(\tau)} - Y_t$ are formed recursively using estimates of equations (12) and (13). These forecasts are then used to form the indicated test statistics, defined in Section 3. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.

3. In each Monte Carlo replication, the simulated test statistics are compared against asymptotic and bootstrapped critical values, using a significance level of 10%. Sections 3.3 and 4.2 describe the asymptotic and bootstrap procedures.

4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 999.

Table 6: Tests of Predictive Power of Capacity Utilization for Inflation
1976:Q1–2004:Q3

	$\tau = 1$	$\tau = 2$	$\tau = 4$	$\tau = 8$	$\tau = 12$
	Summary statistics				
RMSE 1	1.53	1.41	1.41	1.71	1.97
RMSE 2	1.46	1.36	1.31	1.48	1.64
	Test statistics (asymptotic p-values, bootstrap p-values)				
MSE-F	10.92 (.00, .00)	8.47 (.01, .01)	17.69 (.01, .01)	36.18 (.01, .00)	46.21 (.01, .01)
MSE-T	1.02 (.02, .04)	.50 (.07, .10)	.59 (.06, .09)	.98 (.03, .05)	1.13 (.01, .06)
ENC-F	11.77 (.00, .00)	14.66 (.00, .00)	30.86 (.00, .00)	42.88 (.01, .01)	41.91 (.02, .01)
ENC-T	2.09 (.01, .02)	1.72 (.02, .04)	1.92 (.01, .03)	1.75 (.02, .06)	1.56 (.03, .10)

Notes:
1. As described in section 5, forecasts of the τ-period ahead change in inflation ($Y_{t+\tau}^{(\tau)} - Y_t$) are formed recursively using estimates of the restricted model (12) and the unrestricted model (13). Inflation is measured in annualized percentage points. The recursive forecasts are then used to form the indicated test statistics, defined in Section 3.
2. *RMSE 1* and *RMSE 2* refer to the RMSEs of the restricted and unrestricted models (equations (12) and (13)), respectively.
3. The p-values reported in the table are computed with the asymptotic and bootstrap procedures described in sections 3.3 and 4.2.