Tests of Equal Forecast Accuracy for Overlapping Models

Tests of Equal Forecast Accuracy for Overlapping Models *

Todd E. Clark
Federal Reserve Bank of Cleveland

Michael W. McCracken
Federal Reserve Bank of St. Louis

April 2012

Abstract

This paper examines the asymptotic and finite-sample properties of tests of equal forecast accuracy when the models being compared are overlapping in the sense of Vuong (1989). Two models are overlapping when the true model contains just a subset of variables common to the larger sets of variables included in the competing forecasting models. We consider an out-of-sample version of the two-step testing procedure recommended by Vuong but also show that an exact one-step procedure is sometimes applicable. When the models are overlapping, we provide a simple-to-use fixed regressor wild bootstrap that can be used to conduct valid inference. Monte Carlo simulations generally support the theoretical results: the two-step procedure is conservative while the one-step procedure can be accurately sized when appropriate. We conclude with an empirical application comparing the predictive content of credit spreads to growth in real stock prices for forecasting U.S. real GDP growth.

JEL Nos.: C53, C12, C52
Keywords: overlapping models, prediction, out-of-sample

* *Clark*: Economic Research Dept.; Federal Reserve Bank of Cleveland; P.O. Box 6387; Cleveland, OH 44101; todd.clark@clev.frb.org. *McCracken* (corresponding author): Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland, Federal Reserve Bank of St. Louis, Federal Reserve System, or any of its staff.

1 Introduction

Researchers often compare forecasts made by different models to determine whether one model is significantly more accurate than another. Recent examples that compare the predictive content of non-nested models include Corradi, Swanson and Olivetti (2001), Rich et al. (2005), Rapach and Wohar (2007), Naes, Skjeltorp, and Ødegaard (2011), and Fornari and Mele (2011). In each of these papers the authors use what is commonly referred to as a Diebold-Mariano (1995) statistic for equal mean square error (MSE). West (1996) showed the asymptotic distribution of the test applied to forecasts from estimated models to be asymptotically standard normal. Hence, conducting inference is straightforward.

In our previous work on nested model comparisons, we showed that this statistic is typically not asymptotically standard normal and in fact has an asymptotic distribution that has a representation as a function of stochastic integrals of quadratics of Brownian motion.[1] While inference is made much harder than just using normal critical values, in certain instances simulated critical values are available (McCracken 2007). For general conditions, Clark and McCracken (2011) provide a simple to use bootstrap that provides asymptotically valid critical values. Recent examples that compare the predictive content of nested models using the Diebold-Mariano statistic include Hong and Lee (2003), Faust, Rogers, and Wright (2005), Wright and Zhou (2009), and Wegener, von Nitzsch, and Cengiz (2010).

One type of model comparison that does not seem to have penetrated this literature is the comparison of overlapping models. Overlapping models is a concept introduced in Vuong (1989) in the context of comparing the relative fit of two (possibly) misspecified likelihood functions. To get a feel for the problem we address in this paper, let's abstract from likelihood functions and focus on two linear regression functions that are intended to forecast quarterly GDP growth:

$$\begin{aligned} \Delta \log \text{GDP}_{t+1} &= \alpha_{sp500} + \beta_{sp500} sp500_t + \varepsilon_{sp500,t+1} \\ \Delta \log \text{GDP}_{t+1} &= \alpha_{sprd} + \beta_{sprd} sprd_t + \varepsilon_{sprd,t+1}, \end{aligned}$$

where $sp500$ and $sprd$ denote growth in stock prices and a credit spread, respectively.

[1] This result obtains under large R, large P asymptotics, which permit recursive, rolling, and fixed estimation schemes. Giacomini and White (2006) obtain a null asymptotic distribution that is standard normal, under asymptotics that treat R as fixed and P as large, permitting just the rolling and fixed estimation schemes.

As Vuong notes, these two models can have equal predictive content two distinct ways.[2] In the first, both β_{sp500} and β_{sprd} are non-zero and yet $E(\varepsilon^2_{sp500,t+1} - \varepsilon^2_{sprd,t+1}) = E(d_{t+1}) = 0$ despite the fact that $d_{t+1} \equiv \varepsilon^2_{sp500,t+1} - \varepsilon^2_{sprd,t+1} \neq 0$. If this is the case we say the models are non-nested. In the second, both β_{sp500} and β_{sprd} are zero and hence $E(d_{t+1}) = 0$ but in the trivial sense that not only are the two models equally accurate but they are identical in population and hence $\varepsilon_{sp500,t+1} = \varepsilon_{sprd,t+1}$. If this is the case we say the models are overlapping.

As Vuong notes, testing the null hypothesis that the two models are equally accurate (i.e., $E(d_{t+1}) = 0$) becomes much harder when one allows for the possibility that the two models are overlapping. The problem is that the null hypothesis does not uniquely characterize the null asymptotic distribution. If the two models are non-nested the likelihood ratio statistic is asymptotically normal under the null. But if the models are overlapping, the likelihood ratio statistic is mixed chi-square under the null. If one wants to conduct inference with a prespecified type 1 error α, it is not clear which critical values should be used – those from a standard normal or those from a mixed chi-square.

Because of this dichotomy, Vuong suggests a two-step procedure for testing the null hypothesis, using in-sample (as opposed to out-of-sample) statistics. In the first stage, a variance test, conducted at the α_1-percent level, is used to test the null that the population forecast errors are identical and hence the two models are overlapping. If we fail to reject, the procedure stops. Otherwise, if we reject the null (concluding that the two models are not overlapping), we conduct a test of equal accuracy at the α_2-percent level assuming the two models are non-nested. Vuong (1989) argues that this procedure controls the size of the test at the maximum of the nominal sizes used in each stage — i.e., controls $\max(\alpha_1, \alpha_2)$ — and hence the testing procedure is conservative. Subsequent work by Rivers and Vuong (2002) extend Vuong's results in several dimensions, focusing on non-nested models. Most recently, Marcellino and Rossi (2008) generalize the conditions needed to make Vuong's test asymptotically valid, focusing on nested and overlapping models.

[2] There exists a third possibility as well. It can be the case that both β_{sp500} and β_{sprd} are non-zero and $\varepsilon^2_{sp500,t+1} - \varepsilon^2_{sprd,t+1} = 0$. For this to happen it must be the case that $\varepsilon_{sp500,t+1} = \pm\varepsilon_{sprd,t+1}$. While technically feasible, it seems unlikely to be empirically relevant and as such we do not pursue this case throughout the remainder of the paper. In particular, our proposed bootstrap is designed to manage the case in which $\varepsilon^2_{sp500,t+1} - \varepsilon^2_{sprd,t+1} = 0$ but only because β_{sp500} and β_{sprd} are both zero.

In this paper, we extend this previous work on in-sample tests by developing out-of-sample forecast tests of possibly overlapping models. Building on West's (1996) results for non-nested models and Clark and McCracken's (2001, 2005) and McCracken's (2007) results for forecasts from nested models, this paper examines the asymptotic and finite-sample properties of tests of equal forecast accuracy applied to predictions from estimated linear regression models that may be overlapping. We first derive the asymptotic distribution of the Diebold-Mariano-West statistic (that we refer to as the MSE-t statistic) when the models are overlapping. As a corollary to this result, we are also able to derive the asymptotic distribution of the out-of-sample variant of what Vuong (1989) refers to as a variance statistic. When the distribution of the MSE-t statistic is non-normal and one is interested in testing the null of equal accuracy between two overlapping models, we provide a simple-to-use bootstrap that yields asympotically valid critical values. The bootstrap is also applicable for the variance statistic.

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

To assess the practical efficacy of our proposed procedures, we conduct a range of Monte Carlo experiments, and we include an empirical application to forecasts of U.S. GDP growth generated from competing models that could be overlapping. The Monte Carlo analysis shows that the fixed regressor bootstrap developed in this paper has good size and power properties when the models are overlapping under the null hypothesis. Confirming our theoretical work, the simulation results also show our proposed two-step procedure to be conservative and the one-step procedure to be accurately sized when it should be.

The remainder of the paper proceeds as follows. Section 2 introduces the notation, assumptions, and asymptotic results for testing equal accuracy between two overlapping models. Section 3 discusses testing procedures when the models are not known to be overlapping or non-nested. Section 4 presents Monte Carlo results on the performance of our testing procedures, and section 5 applies our tests and bootstrap approach to inference to forecasts of quarterly U.S. real GDP growth. Section 6 concludes.

2 Overlapping Models

We begin by laying out our testing framework when comparing the forecast accuracy of two overlapping models. One of the primary difficulties of working with the overlapping case is that the null hypothesis of equal forecast accuracy, $E(d_{t+1}) = 0$, typically does not uniquely characterize the asymptotic distribution. In one case both models have their own distinct predictive content and hence, for example, β_{sp500} and β_{sprd} are both nonzero. Throughout the remainder of the paper we will assume that when this case arises it is also true that the long-run variance of d_{t+1} is positive and we conclude that the models are non-nested in the sense of West (1996).

In another, the two models degenerate into a baseline model consisting of only those predictors the two models share and hence, for example, $\varepsilon_{sp500,t+1} = \varepsilon_{sprd,t+1}$ because β_{sp500} and β_{sprd} are both zero. This is the case we are interested in within this section. We'll return to the more general situation of inference when one doesn't know which of the two cases holds in section 3.

2.1 Environment

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

For each time t the variable to be predicted is y_{t+1}. The sample is divided into in–sample and out–of–sample portions. The total in–sample observations (on y_t and x_t) span 1 to R. Letting P denote the number of 1–step ahead predictions, the total out-of-sample observations span $R+1$ through $R+P$. The total number of

observations in the sample is $R + P = T$.

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

Both model 1's and model 2's forecasts are generated recursively using estimated parameters. Under this approach both β_1^* and β_2^* are reestimated with added data as forecasting moves forward through time: for $t = R, \ldots, T-1$, model i's ($i = 1, 2$) prediction of y_{t+1} is created using the parameter estimate $\hat{\beta}_{i,t}$ based on data through period t. Models 1 and 2 yield two sequences of P forecast errors, denoted $\hat{u}_{1,t+1} = y_{t+1} - x_{1,t}'\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+1} = y_{t+1} - x_{2,t}'\hat{\beta}_{2,t}$, respectively.

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

$$J_1' = \begin{pmatrix} I_{k_0 \times k_0} & 0_{k_0 \times k_1} \\ 0_{k_1 \times k_0} & I_{k_1 \times k_1} \\ 0_{k_2 \times k_0} & 0_{k_2 \times k_1} \end{pmatrix} \text{ and } J_2' = \begin{pmatrix} I_{k_0 \times k_0} & 0_{k_0 \times k_2} \\ 0_{k_1 \times k_0} & 0_{k_1 \times k_2} \\ 0_{k_2 \times k_0} & I_{k_2 \times k_2} \end{pmatrix} \tag{1}$$

and a $(k_1 + k_2 \times k)$ matrix $\tilde{A}$ satisfying $\tilde{A}'\tilde{A} = B^{-1/2}(-J_1'B_1J_1 + J_2'B_2J_2)B^{-1/2}$, let $\tilde{h}_{t+1} = \sigma^{-1}\tilde{A}B^{1/2}h_{t+1}$, $\tilde{H}(t) = \sigma^{-1}\tilde{A}B^{1/2}H(t)$ and $S_{\tilde{h}\tilde{h}} = E\tilde{h}_{t+1}\tilde{h}_{t+1}'$. Let $W(\omega)$ denote a $(k_1 + k_2 \times 1)$ vector standard Brownian motion.

Given the definitions and forecasting scheme described above, the following assumptions are used to derive the limiting distributions in Theorem 2.1. The assumptions are intended to be only sufficient, not necessary and sufficient.

(A1) The parameters of the forecasting models are estimated using OLS, yielding $\hat{\beta}_{i,t} = \arg\min_{\beta_i} t^{-1}\sum_{s=1}^{t-1}(y_{s+1} - x_{i,s}'\beta_i)^2$, $i = 1, 2$.

(A2) (a) $U_{t+1} = [u_{t+1}, x_t' - Ex_t', h_{t+1}']'$ is covariance stationary. (b) $EU_{t+1} = 0$. (c) $E(h_{t+1}|h_{t+1-j}) = 0$ for $j > 0$. (d) $Ex_tx_t' < \infty$ and is positive definite. (e) For some $r > 8$, U_{t+1} is uniformly L^r bounded. (f) For some $r > d > 2$, U_{t+1} is strong mixing with coefficients of size $-rd/(r-d)$. (g) With $\tilde{U}_{t+1}$ denoting the vector of nonredundant elements of U_{t+1}, $lim_{T\to\infty}T^{-1}E(\sum_{s=1}^{T-1}\tilde{U}_{s+1})(\sum_{s=1}^{T-1}\tilde{U}_{s+1})' = \Omega < \infty$ is

positive definite.

($\underline{A3}$) $\lim_{R,P\to\infty} P/R = \pi \in (0,\infty)$; define $\lambda = (1+\pi)^{-1}$.

($\underline{A3'}$) $\lim_{R,P\to\infty} P/R = 0$; define $\lambda = 1$.

The assumptions provided here are nearly identical to those of Clark and McCracken (2005). We restrict attention to forecasts generated using parameters estimated by OLS (Assumption 1) and we do not allow for processes with either unit roots or time trends (Assumption 2). 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 3) as well as when the in-sample size R is large relative to the out-of-sample size P (Assumption 3′). The assumptions differ only in so far as the notation has changed to accommodate the comparison of overlapping rather than nested models.

2.2 Tests and asymptotic distributions

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal MSE based upon the sequence of loss differentials $\hat{d}_{t+1} = \hat{u}^2_{1,t+1} - \hat{u}^2_{2,t+1}$. If we define $\text{MSE}_i = P^{-1}\sum_{t=R}^{T-1} \hat{u}^2_{i,t+1}$ $(i = 1, 2)$, $\bar{d} = P^{-1}\sum_{t=R}^{T-1} \hat{d}_{t+1} = \text{MSE}_1 - \text{MSE}_2$, and $\hat{S}_{dd} = P^{-1}\sum_{t=R}^{T-1}(\hat{d}_{t+1} - \bar{d})^2$, the statistic takes the form

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

Under the null that both $x_{12,t}$ and $x_{22,t}$ have no predictive power for y_{t+1}, the population difference in MSEs will equal 0. Under the alternative that at least one element of either subvector has predictive power, the population difference in MSEs can be either positive or negative. As a result, when comparing two overlapping models, the MSE-t test is two-sided.

As we will see below, in the case of overlapping models the MSE-t 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_1 and k_2, as well as certain unknown nuisance parameters that depend upon the second moments of the data. In the following define $\Gamma_1 = \int_\lambda^1 \omega^{-1} W(\omega)' S_{\tilde{h}\tilde{h}} dW(\omega)$, $\Gamma_2 = \int_\lambda^1 \omega^{-2} W(\omega)' S_{\tilde{h}\tilde{h}} W(\omega) d\omega$, and $\Gamma_3 = \int_\lambda^1 \omega^{-2} W(\omega)' S^2_{\tilde{h}\tilde{h}} W(\omega) d\omega$. In addition let V_0 and V_1 denote $(k_1 + k_2 \times 1)$ independent standard normal vectors.

Theorem 2.1. (a) Let Assumptions $1-3$ hold. MSE-t $\rightarrow^d (\Gamma_1 - (0.5)\Gamma_2)/\Gamma_3^{1/2}$.
(b) Let Assumptions 1, 2, and $3'$ hold. MSE-t $\rightarrow^d V_0' S_{\tilde{h}\tilde{h}} V_1 / [V_1' S_{\tilde{h}\tilde{h}}^2 V_1]^{1/2} \sim N(0,1)$.

The results in Theorem 2.1 bears a strong resemblance to those discussed in Clark and McCracken (2005). In fact, notationally they are identical. The primary difference is in the definition of the orthogonality condition $\tilde{h}_{t+1}$ and subsequent unknown nuisance parameter $S_{\tilde{h}\tilde{h}}$. Here, $\tilde{h}_{t+1} = \sigma^{-1}\tilde{A}B^{1/2}h_{t+1}$, where $B = (Ex_t x_t')^{-1}$, $h_{t+1} = u_{t+1}x_t$, and $\tilde{A}$ satisfies $\tilde{A}'\tilde{A} = B^{-1/2}(-J_1'B_1J_1 + J_2'B_2J_2)B^{-1/2}$. In the case in which model 2 nests model 1, $\tilde{h}_{t+1} = \sigma^{-1}\tilde{A}B_2^{1/2}h_{2,t+1}$, with $B_i = (Ex_{i,t}x_{i,t}')^{-1}$, $h_{2,t+1} = u_{t+1}x_{2,t}$, and $\tilde{A}$ satisfies $\tilde{A}'\tilde{A} = B_2^{-1/2}(-J'B_1J + B_2)B_2^{-1/2}$, where $J = (I_{k_1\times k_1}, 0_{k_1\times k_2})'$. With such a minor difference in the structure of the problem it is not surprising that the asymptotic distributions are so similar.

Algebraically, the dependence upon $S_{\tilde{h}\tilde{h}}$, which in turn depends upon the second moments of the forecast errors u_{t+1}, the regressors x_t, and the orthogonality conditions h_{t+1}, arises because, in the presence of conditional heteroskedasticity an information matrix-type equality fails. Similarly, in the context of likelihood-ratio statistics, Vuong (1989, Theorem 3.3) shows that the limiting distribution of the likelihood ratio statistic has a representation as a mixture of independent $\chi^2_{(1)}$ variates (in contrast to our integrals of weighted quadratics of Brownian motion). This distribution is free of nuisance parameters when the information matrix equality holds but in general does depend upon such nuisance parameters.

In Theorem 2.1 there are two special cases for which the dependence on $S_{\tilde{h}\tilde{h}}$ is asymptotically irrelevant for the MSE-t statistic. First, 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 distribution no longer depends upon the value of $S_{\tilde{h}\tilde{h}}$. If this is the case we obtain McCracken's (2007) results for MSE-t and thus are able to utilize the estimated asymptotic critical values provided in that paper to conduct inference.[3] Second, in the special case in which $\pi = \lim_{R,P\rightarrow\infty} P/R = 0$, the MSE-$t$ statistic is asymptotically standard normal despite the presence of $S_{\tilde{h}\tilde{h}}$.

Although it may not be immediately apparent, Theorem 2.1 also provides us with the asymptotic distribution of an out-of-sample version of what Vuong referred to

[3]McCracken (2007) provides critical values associated with the upper 90th, 95th and 99th percentiles. Only upper critical values are given because the models were nested. Critical values associated with the lower tail are available upon request from the author.

as a "variance" statistic. Specifically, in the first step of his two-step procedure for testing equal accuracy in the presence of possibly overlapping models, he uses the in-sample likelihood-based version of $P\hat{S}_{dd} = \sum_{t=R}^{T-1}(\hat{d}_{t+1} - \bar{d})^2$ to determine whether or not the two models are overlapping. The following corollary provides the asymptotic distribution for our proposed out-of-sample version of the statistic.

Corollary 2.1. (a) Let Assumptions $1-3$ hold. $P\hat{S}_{dd} \to^d 4\sigma^4\Gamma_3$. (b) Let Assumptions 1, 2, and $3'$ hold. $R\hat{S}_{dd} \to^d 4\sigma^4 V_1' S_{\tilde{h}\tilde{h}}^2 V_1$.[4]

In Corollary 2.1 we find that the asymptotic distribution of the variance statistic takes the form of a stochastic integral when $\lim_{R,P\to\infty} P/R > 0$ but takes the form of a weighted quadratic of vector standard normals when $\lim_{R,P\to\infty} P/R = 0$. The latter result is similar to that in Vuong insofar as the asymptotic distribution is a mixed central chi-square variate. Regardless, in either case (a) or (b), inference is complicated by the presence of the unknown nuisance parameter $S_{\tilde{h}\tilde{h}}$.

2.3 Extensions

For brevity, and to facilitate comparison with the results in Vuong (1989), we have focused attention on methods for inference when the forecasts are one-step ahead and the recursive scheme is used. Most if not all of the results in Theorem 2.1 and Corollary 2.1 can be generalized to situations in which the rolling or fixed schemes are used or when direct multi-step forecasts are made at horizons greater than 1.

Rather than delineate all of these permutations we emphasize two that may prove useful for conducting inference. First, as was the case for nested model comparisons in Clark and McCracken (2005) and McCracken (2007), the recursive, rolling, and fixed schemes are all asymptotically equivalent when $\lim_{R,P\to\infty} P/R = 0$. Stated more precisely, the MSE-t statistic is not only asymptotically standard normal for each of these schemes when $\lim_{R,P\to\infty} P/R = 0$, but also the difference between (say) the MSE-t statistics constructed using the rolling scheme and constructed using the recursive scheme is $o_p(1)$. We immediately conclude that standard normal critical values can be used to conduct inference when $\lim_{R,P\to\infty} P/R = 0$ for each of the schemes when the models are overlapping.

[4]The terms $4\sigma^4$ do not appear in Theorem 2.1 because they cancel with similar terms in the numerator of the MSE-t.

Second, when the fixed scheme is used for model estimation, the MSE-t statistic is asymptotically standard normal regardless of whether assumption 3 or 3′ holds . In particular we find that when the fixed scheme is used to construct forecasts, the MSE-t statistic converges in distribution to $V_0' S_{\tilde{h}\tilde{h}} V_1 / [V_1' S_{\tilde{h}\tilde{h}}^2 V_1]^{1/2} \sim N(0,1)$ regardless of whether $\lim_{R,P\to\infty} P/R$ is zero so long as $\lim_{R,P\to\infty} P/R$ is finite. As described in the next section, this result could be used directly for inference.

3 Testing Procedures

In this section we consider various approaches to testing for equal forecast accuracy when the models may be overlapping. The first two approaches are similar in the sense that they provide a conservative test. The third approach provides accurately sized tests but is only applicable in special instances.

3.1 A conservative two-step procedure

One approach to conducting inference is an out-of-sample version of the two-step procedure suggested by Vuong (1989). In the first stage, $P\hat{S}_{dd}$ (or $R\hat{S}_{dd}$) is used to test whether or not the two models are overlapping at the α_1-percent level. If we fail to reject, the procedure stops. If we reject the null (concluding that the two models are not overlapping) we conduct a test of equal accuracy at the α_2-percent level assuming the two models are non-nested using the MSE-t statistic and standard normal critical values. If we reject, we can conclude that the two models are not equally accurate.

One weakness of this approach is that the critical values for the first step of the procedure are not readily tabulated due to the presence of the unknown nuisance parameter $S_{\tilde{h}\tilde{h}}$. Instead, here we suggest a bootstrap-based approach to estimating asympotically valid critical values. For nested model comparisons, Clark and McCracken (2011) prove that under the null of equal MSE, the resulting critical values are consistent for their population values, while, under the alternative, the test is consistent.[5] Here we delineate the version of the bootstrap appropriate for overlapping models without explicitly proving its asymptotic validity for the comparison of

[5] As shown in Clark and McCracken (2011), the bootstrap is valid with forecast errors that are conditionally heteroskedastic or serially correlated, when the serial correlation arises due to a direct multi-step forecast model setup for horizons greater than one period.

overlapping models. The fixed regressor bootstrap's steps consist of the following.[6]

1. (a) Use OLS to estimate the parameter vector β_0^* associated with the restricted model (the restricted model includes just the variables common to forecasting models 1 and 2). Store the fitted values $x'_{0,s}\hat{\beta}_{0,T}$, $s = 1, ..., T-1$. (b) Use OLS to estimate the parameter vector β^* associated with the unrestricted model that contains all the regressors. Store the residuals $\hat{v}_{s+1}$, $s = 1, ..., T-1$.

2. Let η_s, $s = 1, ..., T$, denote an $i.i.d$ $N(0,1)$ sequence of simulated random variables. Form a time series of innovations $\widehat{v}^*_{s+1} = \eta_{s+1}\widehat{v}_{s+1}$.

4. Form artificial samples of y^*_{s+1} using the fixed regressor structure, $y^*_{s+1} = x'_{0,s}\hat{\beta}_{0,T} + \widehat{v}^*_{s+1}$.

5. Using the artificial data, construct forecasts and an estimate of the test statistic as if these were the original data.

6. Repeat steps 2-5 a large number of times: $j = 1, ..., N$.

7. For the variance ($P\hat{S}_{dd}$) test, reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100-\alpha)$%-ile of the empirical distribution of the simulated test statistics. For the MSE-t test and a significance level of α, compute the lower and upper tail critical values as the $(100-\alpha/2)$ and $(\alpha/2)$ percentiles of the bootstrap distribution.

Simulation evidence provided in section 4 indicates that the bootstrap works very well in the overlapping model context, providing accurately sized tests under the null while still providing considerable power under the alternative.

3.2 A conservative one-step procedure

In most instances, the MSE-t statistic does not have an asymptotically standard normal distribution when the models are overlapping and hence it does not have critical values that are easily accessible. But suppose we knew that $q_{\alpha/2}$ and $q_{1-\alpha/2}$ were the lower and upper $\alpha/2$-percentiles of the null asymptotic distribution of the MSE-t statistic when the models are overlapping. If we knew the models were

[6] Our use of a normal distribution for randomizing the bootstrap errors follows the wild bootstrap development of Goncalves and Kilian (2004). As they note, several approaches could be used to generate random draws with mean 0 and variance 1. Davidson and Flachaire (2008) advocate a two-point approach that can yield asymptotic refinements. We ran a few of our experiments with this two-point randomization scheme and obtained qualitatively similar results. We leave for future research a more thorough investigation of whether the two-point approach used by Davidson and Flachaire in a cross-section context could offer gains in a time series context.

overlapping we would conduct an $\alpha\%$ test by rejecting when either MSE-$t < q_{\alpha/2}$ or MSE-$t > q_{1-\alpha/2}$. Suppose instead that we knew the models were non-nested. We would conduct an $\alpha\%$ test by rejecting when either MSE-$t < z_{\alpha/2}$ or MSE-$t > z_{1-\alpha/2}$, where z_α denotes the α quantile of the standard normal distribution. Unfortunately, the null hypothesis does not tell us which set of critical values should be used to achieve an accurately sized test: $q_{\alpha/2}$ and $q_{1-\alpha/2}$ or $z_{\alpha/2}$ and $z_{1-\alpha/2}$.

To avoid this problem one option is to forego the need for an exact test and simply require a conservative test as we did above, using an out-of-sample version of the two-step procedure recommended by Vuong (1989). In our out-of-sample environment it turns out that the conservative two-step procedure can be replaced with a conservative one-step procedure. This reduction in the number of steps occurs because the same test statistic, applied to either the overlapping or non-nested case, happens to be $O_p(1)$. This was not the case for the likelihood-ratio based methods used in Vuong (1989). There, Vuong restricted the perspective of his inference to only considering likelihood ratio statistics. In that environment he showed that the likelihood ratio statistic was $O_p(T^{-1/2})$ when the models were non-nested but was $O_p(T^{-1})$ when the models were overlapping. Clearly, if one happened to correctly guess which case held, one could conduct valid inference. But if one guessed wrong, the critical values used would be off by an order of magnitude.

But since in our case the MSE-t statistic is bounded in probability under either regime, one could simply use the minimum of the two lower quantiles $min(q_{\alpha/2}, z_{\alpha/2})$ as the upper bound of the lower rejection region and the maximum of the two upper quantiles $\max(q_{1-\alpha/2}, z_{1-\alpha/2})$ as the lower bound of the upper rejection region. This approach would not lead to an exact test at the $\alpha\%$ level but it would be guaranteed to yield a conservative test in large samples.[7] A similar one-step approach could have been used in Vuong (1989) had a wider class of statistics been considered.

One weakness of this approach is that we don't actually know the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$. To get around this issue we exploit the bootstrap described in the previous subsection. Simulation evidence provided in section 4 indicates that the boot-

[7] Admittedly, if the one-step procedure fails to yield a rejection, the models are implied to be equally accurate, but the tester can't be sure if the models are overlapping or non-nested. In light of the interest in the two models being tested, the tester may not care. If he or she does care to distinguish the hypotheses, it would be possible to test each model against the specification that both models nest, using the nested models tests of such studies as Clark and McCracken (2001, 2005, 2011).

strap works very well in this overlapping context, providing accurate estimates of the relevant quantiles under the null while still providing considerable power under the alternative.

3.3 An exact one-step procedure

As we noted in the introduction, there are very special cases in which the MSE-t statistic is asymptotically standard normal when the models are overlapping. If either $\lim_{P,R\to\infty} P/R = 0$ or the fixed scheme is used to construct forecasts, it is not only the case that the MSE-t test is standard normal when the models are overlapping, it is also the case that the test is standard normal when the models are non-nested. Hence, regardless of whether the models are overlapping or non-nested, standard normal critical values can be used to conduct accurately sized inference without having to use a conservative test like those described above.

We should note, however, that this approach is not without its drawbacks. First, in any finite sample it is never the case that $P/R = 0$, and hence it is not obvious how well the standard normal approximation will work when the forecasts are constructed using the recursive scheme and the models are overlapping. It may be the case that the actual size of the test is far from its nominal size and more importantly may not be conservative — one advantage that the other two procedures have. Second, while it is convenient to use standard normal critical values to conduct inference in a one-step procedure, that is hardly a reason to justify using the fixed scheme under the conditions given in this paper. Recall that in our assumptions we essentially assume that the observables are covariance stationary. In this environment one would likely want to use the recursive scheme to evaluate the accuracy of models, since the recursive scheme updates the parameter estimates at each forecast origin as more information is gathered and likely leads to more accurate forecasts.

4 Monte Carlo Evidence

To evaluate the finite sample properties of the above approaches to testing for equal accuracy of forecasts from models that may be overlapping, we use simulations of multivariate data-generating processes (DGPs) with features of common macroeconomic applications. In these simulations, two competing forecasting models include a

common set of variables and another set of variables unique to each model. The null hypothesis is that the forecasts are equally accurate. This null will be satisfied if the models are overlapping or if the models are non-nested. The alternative hypothesis is that one forecast is more accurate than the other.

In the interest of brevity, we focus on forecasts generated under a recursive estimation scheme. However, we have verified that using a rolling estimation scheme yields very similar results. As to testing approaches, we focus on the efficacy of the one-step and two-step testing approaches described above. In light of the validity of the MSE-t test compared against standard normal critical values under certain circumstances or against fixed regressor bootstrap critical values in other circumstances, we also include results for these tests. To help interpret the behaviors of the one-step and two-step procedures, we also provide results for the $P\hat{S}_{dd}$ test that is used in the two-step procedure.

We proceed by first detailing the DGPs and then presenting size and power results, for a forecast horizon of 1 period and a nominal size of 10% (results for 5% are qualitatively the same). Our reported results are based on 10,000 Monte Carlo draws and 499 bootstrap replications.

4.1 Monte Carlo design

For all experiment designs, we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP.[8] With monthly and quarterly data in mind, we consider a range of sample sizes (R, P), for all possible combinations of $R = 50$, 100, 200, and 400 and $P = 20$, 50, 100, and 200.

All experiments use the following general DGP, based loosely on the empirical properties of GDP growth (corresponding to the predictand y), the spread between 10-year and 1-year yields, and the spread between AAA and BBB corporate bonds:

$$\begin{aligned} y_{t+1} &= 0.3y_t + b_1 x_{1,t} + b_2 x_{2,t} + u_{t+1} \\ x_{i,t} &= 0.9x_{i,t-1} + e_{i,t}, \quad i = 1, 2 \\ \text{var}\begin{pmatrix} u_t \\ e_{1,t} \\ e_{2,t} \end{pmatrix} &= \begin{pmatrix} 10.0 & & \\ 0.0 & 0.1 & \\ 0.0 & 0.0 & 0.1 \end{pmatrix}. \end{aligned} \tag{3}$$

[8] The initial observations necessitated by the lag structure of each DGP are generated with draws from the unconditional normal distribution implied by the DGP.

In all experiments, the competing forecasting models are:

$$\text{model 1: } y_{t+1} = \alpha_0 + \alpha_1 y_t + \alpha_2 x_{1,t} + \alpha_3 x_{1,t-1} + u_{1,t+1} \quad (4)$$

$$\text{model 2: } y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 x_{2,t} + \beta_3 x_{2,t-1} + u_{2,t+1}. \quad (5)$$

We consider four different sets of experiments based on this DGP. In the first set, the competing forecasting models are truly overlapping and equally accurate: we parameterize the DGP with $b_1 = b_2 = 0$. In the second set, the competing models are equally accurate, but non-nested instead of overlapping: we use $b_1 = b_2 =$ -2.0 in the DGP. In the third set of experiments, the competing models are also equally accurate and non-nested, but with smaller coefficients on the x variables, such that the variance-based test $P\hat{S}_{dd}$ has less power than in the second set of experiments: in this case, we set $b_1 = b_2 =$ -0.7 in the DGP. In the fourth set of experiments, model 1 is more accurate than model 2, with model 1 corresponding to the DGP: we use b_1 = -2.0 and $b_2 = 0.0$.

4.2 Results

4.2.1 Truly overlapping models

The results in Table 1 indicate that, when the models are truly overlapping, our proposed one- and two-step procedures behave as might be expected under the asymptotic logic described above. The one-step procedure is conservative, yielding an empirical rejection rate that is slightly to somewhat below the nominal size of 10 percent (where nominal refers to the use of 10% critical values from both the standard normal and bootstrap distributions). This one-step test is less undersized when P is small than when P is large. For example, with $R = 100$, the one-step rejection rate is 8.7%, 7.4%, 5.4%, and 3.6% with $P = 20$, 50, 100, and 200, respectively.

The behavior of the one-step approach reflects the properties of individual tests based on comparing the MSE-t statistic against standard normal and bootstrap critical values. Using bootstrap critical values with the MSE-t statistic consistently yields a rejection rate of about 10%. Using standard normal critical values yields a rejection rate that is slightly to somewhat below the intended rate of 10%. Consistent with our asymptotics that imply the MSE-t test to have a standard normal distribution when P/R is close to 0, this testing approach yields a rejection rate closer to 10% when P is small relative to R than when it is large. For example, with $R = 200$, comparing

MSE-t against normal critical values yields empirical size of 11.7% when $P = 20$ and 5.3% when $P = 200$.

In these experiments in which the models are truly overlapping, the two-step procedure is clearly more conservative than the one-step test. Across experiments, the rejection rate is consistently close to 0.1%, which is the product of the 10% nominal sizes of the two tests that enter the two-step approach. The first test in the procedure, the variance-based test $P\hat{S}_{dd}$, is generally correctly sized, except when R and P are both small. As noted above, across all Monte Carlo draws (not just those yielding a rejection by the variance test) the second test in the procedure, MSE-t compared to standard normal critical values, is slightly to modestly undersized. Applying the MSE-t test to the roughly 10 percent of draws in which the variance test yields a rejection leads to rejection in about 10 percent of those draws, for an overall rejection rate of about 1 percent of total draws.

4.2.2 Truly non-nested models

The results in Table 2 show that, when the competing forecasting models are in truth non-nested and equally accurate, our proposed one- and two-step testing procedures behave as intended.

When the coefficients on the x variables are large, as in panel A's experiments, both the one-step and two-step procedures are correctly sized to slightly oversized, yielding a rejection rate of 10 percent or a bit more. For example, with $R = P = 200$, both procedures yield a rejection rate of 10.6%; with $R = 100$ and $P = 50$, the one- and two-step rejection rates are, respectively, 11.9& and 12.1%. The behavior of the two-step test reflects the very high power of the $P\hat{S}_{dd}$ test and an MSE-t test that is modestly oversized (except for large P) when compared against standard normal critical values. Finally, note that comparing the MSE-t test against critical values from the fixed regressor bootstrap yields a rejection rate of more than 10%, as should be expected given that the forecasting models are non-nested, not overlapping.

When the DGP's coefficients on the x variables are non-zero but not large, as in panel B's experiments, our proposed one-step procedure is about correctly sized in experiments with small P and slightly to modestly oversized with large P. The two-step testing approach is generally more conservative than the one-step approach, except when R and P are both large. For example, with $R = 100$ and $P = 20$, the

one-step and two-step rejection rates are 9.9 and 4.4%, respectively. With $R = 100$ and $P = 200$, the corresponding rejection rates are 8.9 and 6.7%. With $R = 400$ and $P = 200$, the one-step and two-step rejection rates are both 9.9%. Again, the behavior of the two-step test is driven in part by the power of the $P\hat{S}_{dd}$ test, which, in most sample size settings, is not as high as in the corresponding experiments with large coefficients on the x variables. Note, though, that these experiments show that the power of the $P\hat{S}_{dd}$ test rises with both R and P.

In these experiments, the behavior of the MSE-t test compared against standard normal critical values across sample sizes doesn't seem to entirely square with asymptotics that imply the test to have a standard normal distribution when P is small relative to R. In our experiments, when P is small relative to R, using standard normal critical values yields a slightly oversized test. For example, in experiments with large coefficients on the x variables and with $P = 20$, comparing the MSE-t test against standard normal critical values yields a rejection rate of 12.8% and 12.5% when $R = 200$ and 400, respectively. It may be that both P and R need to be larger for the small P/R asymptotics to kick in.

4.2.3 Model 1 more accurate than model 2

The results in Table 3 indicate that our proposed one- and two-step testing procedures have comparable power when, in truth, one model is more accurate than another. For example, with $R = 50$ and $P = 100$, the one-step and two-step tests yield rejection rates of 56.2% and 52.5%, respectively. Power rises with both R and P. For instance, with $R = 50$ and $P = 200$, both procedures have power of about 86%. Again, the behavior of these procedures reflects the properties of the component tests, including the $P\hat{S}_{dd}$ test and the MSE-t test compared against standard normal and bootstrap critical values. In these experiments, the $P\hat{S}_{dd}$ test has high power, and the power of the MSE-t test based on the normal and bootstrap critical values is comparable to its power in the one-step and two-step procedures.

4.2.4 Summary of findings

Based on these results, we can recommend both the 1-step and 2-step testing approaches. When the models are overlapping and equally accurate for forecasting, these approaches will reject at a rate at or below nominal size. When the models are

non-nested and equally accurate, these two approaches perform comparably. Overall, consistent with the theory, the two-step procedure is conservative while the one-step procedure can be accurately sized when appropriate. Finally, when one model is more accurate than the other, the tests have similar power.

5 Application

In this section we illustrate the use of our proposed testing approaches with an application to forecasts of quarterly U.S. real GDP growth. With model 1, we forecast GDP growth with a constant, one lag of GDP growth, and two lags of a credit spread defined as the BBB corporate bond rate (from Moody's) less the AAA rate.[9] Model 2 replaces the lags of the credit spread with two lags of growth in real stock prices, where the real stock price is the S&P500 index divided by the price index for personal consumption expenditures less food and energy.[10] We consider one-quarter ahead forecasts for 1985:Q1-2010:Q4. The forecasting models are estimated recursively, with an estimation sample that starts with 1960:Q1.

The results, provided in Table 4, indicate that the two models yield quite similar MSEs. Under our two-step testing procedure for testing equal accuracy, we first compare the test based on the variance of the loss differential to critical values obtained with the fixed regressor bootstrap. For simplicity, we use the simple variance $\hat{S}_{dd}$ rather than the scaled version $P\hat{S}_{dd}$ that has a non-degenerate asymptotic distribution; ignoring the scaling has no effect on the inferences drawn under the bootstrap. In this application, the $\hat{S}_{dd}$ statistic takes the value of 23.664, which exceeds the 90% critical value, but not the 95% critical value. At the 5% significance level, we cannot reject based on the variance statistic; we conclude the models to be overlapping and equally accurate for forecasting.

If we proceed with inference at the 10% significance level, we reject the null that the models are overlapping. Having rejected at the first stage of the test, we proceed to the second stage, of comparing the MSE-t statistic against standard normal critical values, as appropriate with non-nested models. With the MSE-t test taking the value of -0.165, it falls far short of normal critical values at a 10 percent confidence level,

[9] GDP growth is computed as 400 times the log difference of GDP.

[10] We obtained all the data for this application from the FAME database of the Federal Reserve Board of Governors.

so we cannot reject the null of equal accuracy.

Our one-step procedure also implies the two sets of forecasts to be equally accurate. Under this approach, we compare the t-test for equal MSE to the critical values shown in the last two rows of the table. The lower tail critical value is the minimum of the lower tail critical values from the standard normal and bootstrap distributions; the upper tail critical value is the maximum of the upper tail critical values from the standard normal and bootstrap distributions. Since the MSE-t statistic is not close to these critical values, we cannot reject the null of equal forecast accuracy.

6 Conclusion

This paper extends our previous results on nested model comparisons to comparisons of models that may be overlapping. As was the case for Vuong (1989), the main difficulty is handling the fact that the null of equal accuracy does not uniquely characterize the asymptotic distribution of our test statistic, a t-test for equal MSE. In those cases where the models are truly overlapping, the asymptotic distribution is typically non-standard and involves unknown nuisance parameters. A simple-to-use bootstrap is recommended for conducting inference. If it is unknown whether the two models are overlapping, we show how to conduct inference using a conservative two-step procedure akin to that in Vuong (1989). In addition, we show that in certain circumstances an exact one-step procedure is asymptotically valid.

We then conduct a range of Monte Carlo simulations to examine the finite-sample properties of the tests. These experiments indicate our proposed bootstrap has good size and power properties when the models are overlapping under the null. The results generally support the theoretical results: the two step procedure is conservative while the one-step procedure is accurately sized when applicable. In the final part of our analysis, we illustrate the use of our tests with the comparison of forecasting models of real GDP growth.

7 Appendix: Proofs

After adjusting for the change in the definition in $\tilde{h}_{t+1}$ and $S_{\tilde{h}\tilde{h}}$, the proofs are nearly identical to those for nested model comparisons in Clark and McCracken (2005). As such we provide only a sketch of the proofs, noting differences where relevant.

Proof of Theorem 2.1: (a) Maintain Assumption 3 and first consider the numerator of the MSE-t statistic. Extensive algebra and the definition of $\tilde{h}_{t+1}$ imply that $\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) = 2\sigma^2 \sum_{t=R}^{T-1} \tilde{H}'(t)\tilde{h}_{t+1} - \sigma^2 T^{-1} \sum_{t=R}^{T-1}(T^{1/2}\tilde{H}'(t))(T^{1/2}\tilde{H}'(t)) + o_p(1)$. Given Assumption 2, Corollary 29.19 of Davidson (1994) implies $T^{1/2}\tilde{H}(t) \Rightarrow \omega^{-1} S_{\tilde{h}\tilde{h}}^{1/2} W(\omega)$. The Continuous Mapping Theorem and Theorem 3.1 of Hansen (1992) then imply $T^{-1}\sum_{t=R}^{T-1}(T^{1/2}\tilde{H}'(t))(T^{1/2}\tilde{H}'(t)) \rightarrow^d \Gamma_2$ and $\sum_{t=R}^{T-1}\tilde{H}'(t)\tilde{h}_{t+1} \rightarrow^d \Gamma_1$ and hence $\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) \rightarrow^d 2\sigma^2\Gamma_1 - \sigma^2\Gamma_2$.

Now consider the denominator of the MSE-t statistic. Extensive algebra and the definition of $\tilde{h}_{t+1}$ imply that $P\hat{S}_{dd} = 4\sigma^4 \sum_{t=R}^{T-1} \tilde{H}'(t)[E\tilde{h}_{t+1}\tilde{h}_{t+1}']\tilde{H}(t) + o_p(1)$. 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

$$T^{-1}\sum_{t=R}^{T-1} T^{1/2}\tilde{H}'(t) \otimes T^{1/2}\tilde{H}'(t) \rightarrow^d \int_\lambda^1 \omega^{-2}[W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2} \otimes W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2}]d\omega.$$

Since $(\int_\lambda^1 \omega^{-2}[W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2} \otimes W'(\omega)S_{\tilde{h}\tilde{h}}^{1/2}]d\omega)vec[S_{\tilde{h}\tilde{h}}] = \Gamma_3$, we obtain the desired result.

(b) Maintain Assumption 3′ and first consider the numerator of the MSE-t statistic. Extensive algebra and the definition of $\tilde{h}_{2,t+1}$ imply that $\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) = 2\sigma^2(P/R)^{1/2}[R^{1/2}\tilde{H}'(R)][P^{-1/2}\sum_{t=R}^{T-1}\tilde{h}_{t+1}] + o_p((P/R)^{1/2})$. Given Assumption 2, Corollary 29.19 of Davidson (1994) suffices for $(P^{-1/2}\sum_{t=R}^{T-1}\tilde{h}_{t+1}', R^{1/2}\tilde{H}'(R))' \rightarrow^d (V_0' S_{\tilde{h}\tilde{h}}^{1/2}, V_1 S_{\tilde{h}\tilde{h}}^{1/2})'$ for independent $(k_1 + k_2 \times 1)$ standard normal vectors V_0 and V_1 and hence

$$(P/R)^{-1/2}\sum_{t=R}^{T-1}(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2) \rightarrow^d 2\sigma^2 V_0' S_{\tilde{h}\tilde{h}} V_1.$$

Now consider the denominator of the MSE-t statistic. Extensive algebra and the definition of $\tilde{h}_{t+1}$ imply that $P\hat{S}_{dd} = 4(P/R)\sigma^4[R^{1/2}\tilde{H}'(R)][E\tilde{h}_{t+1}\tilde{h}_{t+1}'][R^{1/2}\tilde{H}(R)] + o_p(P/R)$. Since $R^{1/2}\tilde{H}(R) \rightarrow^d S_{\tilde{h}\tilde{h}}^{1/2} V_1$ we immediately find that $(P/R)^{-1}P\hat{S}_{dd} \rightarrow^d 4\sigma^4 V_1' S_{\tilde{h}\tilde{h}}^2 V_1$ and we obtain the desired result.

Proof of Corollary 2.1: Follows immediately from the proof of Theorem 2.1.

References

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

Clark, T. E., McCracken, M. W. (2005). Evaluating direct multi-step forecasts. *Econometric Reviews* 24:369-404.

Clark, T. E., McCracken, M. W. (2011). Reality checks and comparisons of nested predictive models. *Journal of Business and Economic Statistics*, forthcoming.

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

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

Davidson, R., Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics* 146:162-169.

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

Faust, J., Rogers, J.H., Wright, J. (2005). News and noise in G-7 GDP announcements. *Journal of Money, Credit and Banking* 37:403-19.

Fornari, F., Mele, A. (2011). Financial volatility and economic activity. Manuscript, LSE and ECB.

Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica* 74:1545-1578.

Hansen, B. E. (1992). Convergence to stochastic integrals for dependent heterogeneous processes. *Econometric Theory* 8:489-500.

Hong, Y., Lee, T. H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85:1048-1062.

Marcellino, M., Rossi, B. (2008). Model selection for nested and overlapping nonlinear, dynamic and possibly mis-specified models. *Oxford Bulletin of Economics and Statistics* 70:867-893.

McCracken, M. W. (2007). Asymptotics for out–of–sample tests of causality. *Journal of Econometrics* 140:719-752.

Naes, R., Skjeltorp, J. A., Ødegaard, B. A. (2011). Stock market liquidity and the business cycle. *Journal of Finance* 66:139-176.

Rapach, D., Wohar, M. (2007). Forecasting the recent behavior of US business fixed

investment spending: an analysis of competing models. *Journal of Forecasting* 26:33-51.

Rich, R., Bram, J., Haughwout A., Orr, J., Rosen, R., Sela, R. (2005). Using regional economic indexes to forecast tax bases: Evidence from New York. *Review of Economics and Statistics* 87:627-634.

Rivers, D., Vuong, Q. (2002). Model selection tests for nonlinear dynamics models. *Econometrics Journal* 5:1-39.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57:307-333.

Wegener, C., von Nitzscha, R., Cengiz, C. (2010). An advanced perspective on the predictability in hedge fund returns. *Journal of Banking and Finance* 34:2694-2708.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica* 64:1067-1084.

Wright, J., Zhou, H. (2009). Bond risk premia and realized jump risk. *Journal of Banking and Finance* 33:2333-2345.

Table 1: Monte Carlo Rejection Rates: Overlapping Models

(nominal size = 10%)

R	P	two-step test	one-step test	MSE-t vs. N(0,1)	MSE-t vs. bootstrap	$P\hat{S}_{dd}$
50	20	0.015	0.081	0.104	0.098	0.164
50	50	0.009	0.059	0.068	0.091	0.146
50	100	0.006	0.039	0.042	0.091	0.128
50	200	0.002	0.023	0.023	0.088	0.113
100	20	0.011	0.087	0.109	0.099	0.129
100	50	0.010	0.074	0.084	0.099	0.126
100	100	0.007	0.054	0.057	0.094	0.120
100	200	0.005	0.036	0.037	0.092	0.113
200	20	0.010	0.095	0.117	0.101	0.104
200	50	0.010	0.089	0.097	0.102	0.106
200	100	0.009	0.072	0.077	0.100	0.109
200	200	0.007	0.053	0.053	0.098	0.108
400	20	0.008	0.100	0.120	0.105	0.091
400	50	0.008	0.093	0.102	0.102	0.097
400	100	0.009	0.085	0.092	0.104	0.099
400	200	0.007	0.064	0.066	0.099	0.098

Notes:
1. The data generating process is defined in equation (3). In these experiments, the coefficients $b_1 = b_2 = 0$, such that the competing forecasting models are overlapping and equally accurate in forecasting.
2. For each artificial data set, forecasts of y_{t+1} are formed recursively using estimates of equations (4) and (5). These forecasts are then used to form the indicated test statistics. R and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively).
3. The test statistics MSE-t and $P\hat{S}_{dd}$ are defined in section 2.2. The fixed regressor bootstrap and the one- and two-step procedures are defined in sections 3.1 and 3.2.
4. The number of Monte Carlo simulations is 10,000; the number of bootstrap draws is 499.

Table 2: Monte Carlo Rejection Rates: Non-Nested Models

(nominal size = 10%)

R	P	two-step test	one-step test	MSE-t vs. N(0,1)	MSE-t vs. bootstrap	$P\hat{S}_{dd}$
A. Large x coefficients: $b_1 = b_2 = -2.0$						
50	20	0.103	0.117	0.132	0.153	0.831
50	50	0.113	0.116	0.121	0.199	0.940
50	100	0.110	0.111	0.113	0.230	0.988
50	200	0.107	0.106	0.107	0.263	0.999
100	20	0.118	0.114	0.130	0.134	0.942
100	50	0.121	0.119	0.123	0.170	0.989
100	100	0.117	0.115	0.117	0.194	0.999
100	200	0.124	0.124	0.124	0.239	1.000
200	20	0.125	0.113	0.128	0.122	0.990
200	50	0.117	0.113	0.117	0.140	1.000
200	100	0.124	0.123	0.124	0.168	1.000
200	200	0.106	0.106	0.106	0.183	1.000
400	20	0.125	0.112	0.125	0.117	0.999
400	50	0.119	0.113	0.119	0.127	1.000
400	100	0.124	0.122	0.124	0.146	1.000
400	200	0.117	0.116	0.117	0.165	1.000
B. Small x coefficients: $b_1 = b_2 = -0.7$						
50	20	0.032	0.089	0.115	0.114	0.338
50	50	0.031	0.082	0.090	0.132	0.387
50	100	0.033	0.073	0.075	0.154	0.468
50	200	0.040	0.068	0.068	0.198	0.610
100	20	0.044	0.099	0.121	0.115	0.422
100	50	0.048	0.094	0.102	0.129	0.510
100	100	0.051	0.089	0.091	0.152	0.612
100	200	0.067	0.089	0.090	0.193	0.757
200	20	0.065	0.108	0.126	0.117	0.586
200	50	0.071	0.099	0.107	0.122	0.709
200	100	0.087	0.106	0.109	0.146	0.811
200	200	0.084	0.094	0.095	0.160	0.902
400	20	0.096	0.114	0.131	0.119	0.804
400	50	0.100	0.105	0.112	0.120	0.908
400	100	0.106	0.107	0.111	0.133	0.959
400	200	0.099	0.099	0.101	0.142	0.987

Notes:
1. The data generating process is defined in equation (3). In the experiments in panel A, the DGP coefficients are set to $b_1 = b_2 = -2.0$. In the experiments in panel B, the DGP coefficients are set to $b_1 = b_2 = -0.7$.
2. See the notes to Table 1.

Table 3: Monte Carlo Rejection Rates: Model 1 Is DGP

(nominal size = 10%)

R	P	two-step test	one-step test	MSE-t vs. N(0,1)	MSE-t vs. bootstrap	$P\hat{S}_{dd}$
50	20	0.132	0.167	0.197	0.204	0.670
50	50	0.264	0.306	0.320	0.402	0.797
50	100	0.525	0.562	0.567	0.707	0.911
50	200	0.860	0.869	0.869	0.954	0.987
100	20	0.185	0.191	0.222	0.209	0.838
100	50	0.356	0.358	0.374	0.420	0.940
100	100	0.595	0.599	0.605	0.694	0.983
100	200	0.883	0.884	0.884	0.945	0.999
200	20	0.216	0.199	0.228	0.209	0.958
200	50	0.391	0.376	0.393	0.408	0.994
200	100	0.620	0.614	0.620	0.675	1.000
200	200	0.892	0.891	0.892	0.932	1.000
400	20	0.237	0.208	0.239	0.215	0.995
400	50	0.405	0.384	0.405	0.403	1.000
400	100	0.642	0.632	0.642	0.669	1.000
400	200	0.890	0.889	0.890	0.918	1.000

Notes:
1. The data generating process is defined in equation (3). In these experiments, the DGP coefficients are set to $b_1 = -2.0$, $b_2 = 0$.
2. See the notes to Table 1.

Table 4: Results of Application to Forecasts of GDP Growth

MSE_1	4.900
MSE_2	4.978
$MSE_1 - MSE_2$	-0.078
$\hat{S}_{dd}$	23.664
90%, 95% bootstrap critical values for $\hat{S}_{dd}$	16.376, 24.183
MSE-t	-0.165
90% bootstrap critical values for MSE-t	-1.707, 1.494
95% bootstrap critical values for MSE-t	-1.972, 1.758
1-step procedure 90% critical values for MSE-t	-1.707, 1.645
1-step procedure 95% critical values for MSE-t	-1.972, 1.960

Notes:
1. As described in section 5, 1-quarter ahead forecasts of real GDP growth (defined as 400 times the log difference of real GDP) are generated recursively from competing models including either a credit spread (model 1) or growth in real stock prices (model 2). Forecasts from 1985:Q1 through 2010:Q4 are obtained from models estimated with a data sample starting in 1960:Q1.
2. The test statistics MSE-t and $\hat{S}_{dd}$ are defined in section 2.2. The fixed regressor bootstrap and the one- and two-step procedures are defined in sections 3.1 and 3.2.
3. The number of bootstrap draws is 4999.