Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 138 (2007) 291–311
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Approximately normal tests for equal predictive accuracy in nested models

Todd E. Clark[a,*], Kenneth D. West[b]

[a]*Economic Research Department, Federal Reserve Bank of Kansas City, 925 Grand Blvd., Kansas City, MO 64198, USA*
[b]*University of Wisconsin, WI, USA*

Available online 18 July 2006

Abstract

Forecast evaluation often compares a parsimonious null model to a larger model that nests the null model. Under the null that the parsimonious model generates the data, the larger model introduces noise into its forecasts by estimating parameters whose population values are zero. We observe that the mean squared prediction error (MSPE) from the parsimonious model is therefore expected to be *smaller* than that of the larger model. We describe how to adjust MSPEs to account for this noise. We propose applying standard methods [West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084] to test whether the adjusted mean squared error difference is zero. We refer to nonstandard limiting distributions derived in Clark and McCracken [2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110; 2005a. Evaluating direct multistep forecasts. Econometric Reviews 24, 369–404] to argue that use of standard normal critical values will yield actual sizes close to, but a little less than, nominal size. Simulation evidence supports our recommended procedure.

JEL classification: C220; C530; E170; F370

Keywords: Out of sample; Causality; Random walk; Testing; Efficient markets; Principle of parsimony

*Corresponding author. Tel.: +1 816 881 2575; fax: +1 816 881 2199.
E-mail address: todd.e.clark@kc.frb.org (T.E. Clark).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2006.05.023

1. Introduction

Forecast evaluation in economics often involves a comparison of a parsimonious null model to a larger alternative model that nests the parsimonious model. Such comparisons are common in both asset pricing and macroeconomic applications. In asset-pricing applications, the parsimonious benchmark model usually is one that posits that an expected return is constant. The larger alternative model attempts to use time-varying variables to predict returns. If the asset in question is equities, for example, a possible predictor is the dividend-price ratio. In macroeconomic applications, the parsimonious model might be a univariate autoregression for the variable to be predicted. The larger alternative model might be a bivariate or multivariate vector autoregression (VAR) that includes lags of some variables in addition to lags of the variable to be predicted. If the variable to be predicted is inflation, for example, the VAR might be bivariate and include lags of the output gap along with lags of inflation.

Perhaps the most commonly used statistic for comparisons of predictions from nested models is mean squared prediction error (MSPE).[1] In this paper we explore the behavior of standard normal inference for MSPE in comparisons of nested models.

Our starting point relates to an observation made in our earlier work (Clark and West (2005)): under the null that the additional parameters in the alternative model do not help prediction, the MSPE of the parsimonious model should be *smaller* than that of the alternative. This is true even though the null states that with parameters set at their population values, the larger model reduces to the parsimonious model, implying that the two models have equal MSPE when parameters are set at population values. The intuition for the smaller MSPE for the parsimonious model is that the parsimonious model gains efficiency by setting to zero parameters that are zero in population, while the alternative introduces noise into the forecasting process that will, in finite samples, inflate its MSPE. Our earlier paper (Clark and West, 2005) assumed that the parsimonious model is a random walk. The present paper allows a general parametric specification for the parsimonious model. This complicates the asymptotic theory, though in the end our recommendation for applied researchers is a straightforward generalization of our recommendation in Clark and West (2005).

Specifically, we recommend that the point estimate of the difference between the MSPEs of the two models be adjusted for the noise associated with the larger model's forecast. We describe a simple method to do so. We suggest as well that standard procedures (Diebold and Mariano, 1995; West, 1996) be used to compute a standard error for the MSPE difference adjusted for such noise. As in Clark and West (2005), we call the resulting statistic *MSPE-adjusted*. As has been standard in the literature on comparing forecasts from nested models since the initial paper by Ashley et al. (1980), we consider one-sided tests. The alternative is that the large model has smaller MSPE.

In contrast to the simple Clark and West (2005) environment, under our preferred set of technical conditions the MSPE-adjusted statistic is *not* asymptotically normal. But we refer to the quantiles of a certain non-standard distribution studied in Clark and McCracken

[1] References include Lettau and Ludvigson (2001), Stock and Watson (2002, 2003, 2004), Goyal and Welch (2003), Marcellino et al. (2003), Diebold and Li (2006), Orphanides and van Norden (2005), Rapach and Weber (2004), Clark and McCracken (2005b) and Shintani (2005).

(2001, 2005a) to argue that standard normal critical values will yield actual sizes close to, but a little less than, nominal size, for samples sufficiently large.

Our simulations show that these quantiles are applicable with samples of size typically available. We report results from 48 sets of simulations on one step ahead forecasts, with the sets of simulations varying largely in terms of sample size, but as well in terms of DGP. In all 48 simulations, use of the 0.10 normal critical value of 1.282 resulted in actual size between 0.05 and 0.10. The median size across the 48 sets was about 0.08. Forecasts generated using rolling regressions generally yielded more accurately sized tests than those using recursive regressions. Comparable results apply when we use the 0.05 normal critical value of 1.645: the median size is about 0.04. These results are consistent with the simulations in Clark and McCracken (2001, 2005a).

By contrast, standard normal inference for the raw (unadjusted) difference in MSPEs—called "*MSPE-normal*" in our tables—performed abysmally. For one-step ahead forecasts and nominal 0.10 tests, the median size across 48 sets of simulations was less than 0.01, for example. The poor performance is consistent with the asymptotic theory and simulations in McCracken (2004) and Clark and McCracken (2001, 2005a).

Of course, one might use simulation-based methods to conduct inference on MSPE-adjusted, or, for that matter, MSPE-normal. One such method would be a bootstrap, applied in forecasting contexts by Mark (1995), Kilian (1999), Clark and West (2005), and Clark and McCracken (2005a). Our simulations find that the bootstrap results in a modest improvement relative to MSPE-adjusted, with a median size across 48 sets of simulation between 0.09 and 0.10. Another simulation method we examine is to simulate the non–standard limiting distributions of the tests, as in Clark and McCracken (2005a). We find that such a simulation-based method also results in modest improvements in size relative to MSPE-adjusted (median size across 48 sets of simulations about 0.11).

Our simulations also examine a certain statistic for nested models proposed by Chao et al. (2001) ("*CCS*", in our tables).[2] We find CCS performs a little better than does MSPE-adjusted in terms of size, somewhat more poorly in terms of power. (By construction, size adjusted power is identical for MSPE-adjusted and for the simulation-based methods described in the previous paragraph.) A not-for-publication appendix reports results for multistep forecasts for a subset of the DGPs reported in our tables. We find that on balance, the bootstrap performs distinctly better than MSPE-adjusted for relatively small samples sizes, comparably for medium or larger sample sizes; overall, MSPE-adjusted performs a little better than CCS, a lot better than MSPE-normal.

We interpret these results as supporting the use of MSPE-adjusted, with standard normal critical values, in forecast comparisons of nested models. MSPE-adjusted allows inference just about as accurate as the other tests we investigate, with power that is as good or better, and with ease of interpretation that empirical researchers find appealing.

Readers uninterested in theoretical or simulation details need only read Section 2, which outlines computation of MSPE-adjusted in what we hope is a self-contained way. Section 3 describes the setup and computation of point estimates. Section 4 describes the theory

[2] A previous version of this paper presented results for an encompassing statistic proposed by Chong and Hendry (1986). This statistic performed quite poorly. A referee has properly noted that it is of interest to consider still other loss functions, including ones based on economic rather than statistical criteria. We defer such analysis to future research.

underlying inference about MSPE-adjusted. Section 5 describes construction of test statistics. Section 6 presents simulation results.

Section 7 presents an empirical example. Section 8 concludes. An Appendix available on request from the authors includes some results omitted from this paper to save space.

2. MSPE-adjusted

We present our recommended procedure using what we hope is self-explanatory notation. Exact definitions are in subsequent sections.

Model 1 is the parsimonious model. Model 2 is the larger model that nests model 1—that is, model 2 reduces to model 1 if some model 2 parameters are set to zero. The researcher is interested in τ-step ahead forecasts. The period t forecasts of $y_{t+\tau}$ from the two models are denoted $\hat{y}_{1t,t+\tau}$ and $\hat{y}_{2t,t+\tau}$ with corresponding period $t+\tau$ forecast errors $y_{t+\tau}-\hat{y}_{1t,t+\tau}$ and $y_{t+\tau}-\hat{y}_{2t,t+\tau}$. The sample MSPEs are $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, computed as sample averages of $(y_{t+\tau}-\hat{y}_{1t,t+\tau})^2$ and $(y_{t+\tau}-\hat{y}_{2t,t+\tau})^2$. Define a term "adj." (as in "adjustment") as the sample average of $(\hat{y}_{1t,t+\tau}-\hat{y}_{2t,t+\tau})^2$. Define $\hat{\sigma}_2^2$-adj. as the difference between $\hat{\sigma}_2^2$ and the "adj." term just defined. Let P be the number of predictions used in computing these averages.

Thus,

$$\hat{\sigma}_1^2 = P^{-1}\sum(y_{t+\tau}-\hat{y}_{1t,t+\tau})^2, \hat{\sigma}_2^2 = P^{-1}\sum(y_{t+\tau}-\hat{y}_{2t,t+\tau})^2,$$
$$\hat{\sigma}_2^2-\text{adj.} = P^{-1}\sum(y_{t+\tau}-\hat{y}_{2t,t+\tau})^2 - P^{-1}\sum(\hat{y}_{1t,t+\tau}-\hat{y}_{2t,t+\tau})^2.$$

The null hypothesis is equal MSPE. The alternative is that model 2 has a smaller MSPE than model 1. We propose testing the null by examining not $\hat{\sigma}_1^2-\hat{\sigma}_2^2$ but $\hat{\sigma}_1^2-(\hat{\sigma}_2^2-\text{adj.})$, rejecting if this difference is sufficiently positive. Note that $(\hat{\sigma}_2^2-\text{adj.})<\hat{\sigma}_2^2$ so the "adj." term adjusts for the upward bias in MSPE produced by estimation of parameters that are zero under the null.

Perhaps the computationally most convenient way to proceed is to define:

$$\hat{f}_{t+\tau} = (y_{t+\tau}-\hat{y}_{1t,t+\tau})^2 - [(y_{t+\tau}-\hat{y}_{2t,t+\tau})^2 - (\hat{y}_{1t,t+\tau}-\hat{y}_{2t,t+\tau})^2]. \tag{2.1}$$

Now, $\hat{\sigma}_1^2-(\hat{\sigma}_2^2-\text{adj.})$ is simply the sample average of $\hat{f}_{t+\tau}$. So test for equal MSPE by regressing $\hat{f}_{t+\tau}$ on a constant and using the resulting t-statistic for a zero coefficient. Reject if this statistic is greater than $+1.282$ (for a one sided 0.10 test) or $+1.645$ (for a one sided 0.05 test). For one step ahead forecast errors, the usual least squares standard error can be used. For autocorrelated forecast errors, an autocorrelation consistent standard error should be used.

3. Environment

Let model 1 be the parsimonious model, model 2 the larger model. Sometimes we will refer to model 1 as the null model, model 2 as the alternative model. For simplicity we assume the models are linear and are estimated by least squares. Computation of test statistics for nonlinear parametric models is straightforward, though certain of our asymptotic results may not generalize, as noted below. Let y_t be a scalar random variable whose prediction is of interest. The parsimonious model uses a vector X_{1t} to predict y_t. The alternative uses a vector X_{2t}, with the elements of X_{1t} a strict subset

of the elements of X_{2t}:

$$\text{Model 1}: \quad y_t = X_{1t}'\beta_1^* + e_{1t}, Ee_{1t}X_{1t} = 0, \tag{3.1}$$

$$\text{Model 2}: \quad y_t = X_{1t}'\delta^* + Z_t'\gamma^* + e_{2t} \equiv X_{2t}'\beta_2^* + e_{2t}, X_{2t}' \equiv (X_{1t}', Z_t')', \quad \beta_2^{*\prime} = (\delta^{*\prime}, \gamma^{*\prime})', Ee_{2t}X_{2t} = 0. \tag{3.2}$$

In (3.1) and (3.2), $E(y_t|X_{1t}) = X_{1t}'\beta_1^*$ and $E(y_t|X_{2t}) = X_{2t}'\beta_2^*$. Of course, $\delta^* = \beta_1^*$ if $EX_{1t}Z_t' = 0$, or if, as discussed below, $\gamma^* = 0$. In (3.1) and (3.2), the unobservable regression disturbances e_{1t} and e_{2t} may be serially correlated. That is, we allow setups where overlapping data are used in forming multistep predictions, in which case the disturbances follow an MA process of whose order is one less than the forecast horizon. As well, the disturbances may be heteroskedastic conditional on the right-hand side (r.h.s.) variables. Our dating presumes that X_{1t} and X_{2t} are observed prior to y_t and so can be used to predict y_t. For example, if the parsimonious model is an AR(1), X_{1t} is bivariate with $X_{1t} = (1, y_{t-1})'$.

As is indicated in (3.2), model 2 nests model 1 in the sense that when $\gamma^* = 0$, model 2 reduces to model 1. So under the null,

$$\gamma^* = 0, \ \beta_2^{*\prime} = (\beta_1^{*\prime} 0'), \ X_{1t}'\beta_1^* = X_{2t}'\beta_2^*, \ e_{1t} = e_{2t} \equiv e_t, \tag{3.3}$$

The implication of (3.3) that we examine is

$$Ee_{1t}^2 - Ee_{2t}^2 = 0 \quad \text{(equal MSPE)}. \tag{3.4}$$

Under the alternative, the additional variables used by model 2 provide additional predictive ability ($\gamma^* \neq 0$):

$$\gamma^* \neq 0, \quad Ee_{1t}^2 - Ee_{2t}^2 > 0, \quad Ee_{1t}Z_t' \neq 0. \tag{3.5}$$

To explain how one uses out of sample prediction errors to test (3.4), assume for simplicity that forecasts are one step ahead, with obvious generalization to multistep forecasts. Let the total sample size be $T+1$. The last P observations of this sample are used for forecast evaluation. The first R observations are used to construct an initial set of regression estimates that are then used for the first prediction. We have $R + P = T + 1$. Let $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$ denote least-squares estimates that rely on data from period t or earlier, constructed using either the rolling or recursive scheme.[3] Asymptotic and finite sample results differ for the two schemes. Examples of applications using each of these schemes include Campbell and Thompson (2005) and Faust et al. (2005) (recursive) and Cooper et al. (2005) and Ang et al. (2004) (rolling).

Write the predictions and prediction errors as

$$\hat{y}_{1t+1} = X_{1t+1}'\hat{\beta}_{1t}, \quad \hat{e}_{1t+1} \equiv y_{t+1} - \hat{y}_{1t+1}, \quad \hat{y}_{2t+1} \equiv X_{2t+1}'\hat{\beta}_{2t}, \quad \hat{e}_{2t+1} \equiv y_{t+1} - \hat{y}_{2t+1}. \tag{3.6}$$

[3] In the recursive scheme, the size of the sample used to estimate β grows as one makes predictions for successive observations. One first estimates β_1^* and β_2^* with data from 1 to R and uses the estimate to predict; one next estimates β_1^* and β_2^* with data from 1 to $R+1$, with the new estimate used to predict, and so on. In the rolling scheme, the sequence of regression estimates is always generated from a sample of size R. The first estimates of β_1^* and β_2^* are obtained with a sample running from 1 to R, the next with a sample running from 2 to $R+1, \ldots$, and so on. See West (2005).

(In the notation of Section 2, $\hat{y}_{1t+1} = \hat{y}_{1t,t+1}$, and $\hat{y}_{2t+1} = \hat{y}_{2t,t+1}$ a simplification of subscripts afforded by our expositional decision to focus in this section on one step ahead forecasts.) Then the sample analog that may be used to test (3.4) is:

$$P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}^2 - P^{-1}\sum_{t=R}^{T}\hat{e}_{2t+1}^2 \equiv \hat{\sigma}_1^2 - \hat{\sigma}_2^2 \quad \text{(MSPE-normal)}. \tag{3.7}$$

The introduction remarked that under the null, we expect the sample MSPE from the parsimonious model to be smaller than that from the alternative model. To illustrate that result, and to motivate that "MSPE-adjusted" statistic that we propose, observe that algebraic manipulations yield:

$$\hat{e}_{1t+1}^2 - \hat{e}_{2t+1}^2 = -2\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1}) - (\hat{y}_{1t+1} - \hat{y}_{2t+1})^2.$$

Thus MSPE-normal may be written

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \equiv P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}^2 - P^{-1}\sum_{t=R}^{T}\hat{e}_{2t+1}^2 = -2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1}) - P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2. \tag{3.8}$$

Under the null, e_{1t} is uncorrelated with both X_{1t} and X_{2t}. It seems reasonable to expect, then, that $P^{-1}\Sigma_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1}) \approx 0$ (though as discussed below not all seemingly reasonable asymptotic approximations imply that a large sample average of $\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1})$ will be zero). Since $-P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2 < 0$, *we expect the sample MSPE from the parsimonious model to be less than that of the alternative model.* The obvious adjustment to properly center the statistic so that it will, under the null, have approximate mean zero, is to adjust for the negative term $-P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$. As in Clark and West (2005), we call this *MSPE-adjusted*:

$$P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}^2 - \left[P^{-1}\sum_{t=R}^{T}\hat{e}_{2t+1}^2 - P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2\right] \equiv \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})$$

(MSPE-adjusted). (3.9)

We see from (3.8) that MSPE-adjusted is

$$\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.}) = -2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1}). \tag{3.10}$$

With a little algebra, it may be shown that under the alternative (3.5), $-Ee_{1t}(X_{1t}'\beta_1^* - X_{2t}'\beta_2^*) > 0$. Thus we expect MSPE-adjusted to be positive, and we use one-tailed tests in our simulations and empirical examples.

We shall compare, via simulations, the performance of t-statistics associated with MSPE-normal (3.7) and MSPE-adjusted (3.9). To our knowledge there is no appealing and general set of conditions under which the t-statistics computed using MSPE-normal are asymptotically normal.[4] The presence of the negative term $-P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$

[4]Standard critical values are appropriate when $P/R \to 0$ under an asymptotic approximation in which $R \to \infty$, $P \to \infty$ (West 1996; West and McCracken (1998); McCracken 2004; Clark and McCracken 2001, 2005a).

causes this statistic to be miscentered. We use standard critical values in part because some practitioners have used such values (e.g., Goyal and Welch, 2003), in part to contrast this t-statistic to that of other statistics. Asymptotic properties of t-statistics for MSPE-adjusted are discussed in the next section.

4. Inference on MSPE-adjusted

With a little algebra, it can be established that

$$\text{MSPE-adjusted} = 2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{e}_{1t+1}-\hat{e}_{2t+1}). \tag{4.1}$$

Harvey et al. (1998) propounded testing $Ee_{1t}(e_{1t}-e_{2t})=0$, arguing that this is an attractive implication of encompassing. Thus one can interpret us as proposing that a comparison of MSPEs be transformed into an encompassing test, though our preferred interpretation is that we are executing a comparison of MSPEs after adjusting for the upward bias in the MSPE of the larger model.[5]

In analysis of (4.1), for the most part we follow Clark and McCracken (2001, 2005a). These papers require that the estimator of regression parameters be nonlinear least squares (ordinary least squares of course a special case). They also require that multistep forecasts be made with what is called the "direct" rather than "iterated" method (see, e.g., Marcellino et al., 2004).

When (4.1) is divided by the usual asymptotic standard error, Clark and McCracken call the result "*Enc-t.*" Their results for Enc-t include the following. When $R\to\infty$, $P\to\infty$ with R/P approaching a finite nonzero constant, Enc-t is Op(1), with a non-standard limiting distribution. This result applies for both one step ahead and multistep ahead forecasts, and for conditionally heteroskedastic as well as conditionally homoskedastic forecast errors.

For one step ahead forecasts in conditionally homoskedastic environments, Clark and McCracken write the limiting distribution of Enc-t as a functional of Brownian motion that does not depend on the specifics of the DGP. The functional does depend on: (a) the difference between the dimension of X_{2t} and X_{1t} (i.e., the dimension of Z_t in (3.2)), (b) the large sample limit of P/R; (c) whether the rolling or recursive scheme is used. In an unpublished appendix to Clark and McCracken (2001) that may be found on Clark's web page (⟨www.kc.frb.org/Econres/staff/tec.htm⟩), quantiles are given for $1\leqslant$ dimension of $Z_t\leqslant 20$ and for 20 different limiting values of P/R, ranging from $P/R=0.1$ to 20.0, with separate tables for rolling and recursive sampling schemes. Upon inspection of 400 sets of quantiles (one set for each value of the dimension of Z_t and each limiting value of P/R), one sees that apart from a couple of exceptions, and for both rolling and

(*footnote continued*)
However, in many applications P is small relative to R but not so small as to make $P/R\to 0$ obviously attractive, an inference supported by simulation results reported below.

[5]Our preferred interpretation permits us to distinguish between tests of $Ee_{1t}(e_{1t}-e_{2t})=0$ in nested and nonnested models. We are about to argue that in nested models, conventional standard errors yield an asymptotic normal approximation that is accurate for practical purposes. West's (2001) simulations illustrate that in nonnested models, conventional standard errors can lead to seriously misleading inference. Incidentally, in the sample sizes we consider, the degrees of freedom and critical value (t rather than normal) adjustments suggested by Harvey et al. (1998) will have negligible effects.

recursive schemes,

$$0.90 \text{ quantile} \leqslant 1.282 \leqslant 0.95 \text{ quantile}. \tag{4.2}$$

Recall that for one-tailed tests using standard normal critical values, the 0.90 quantile is 1.282. The implication is that for P and R sufficiently large — again, the Clark and McCracken (2001) asymptotics require $R \to \infty$ and $P \to \infty$ — *for one step ahead predictions in conditionally homoskedastic environments, standard normal inference on MSPE-adjusted will lead to nominal* 0.10 *tests that have actual size somewhere between* 0.05 *and* 0.10.

We are confident that this implication is one that can be relied on in practice. We stress, however, that we have no formal proof of the claim, nor do we even assert that the italicized assertion is literally true: we consider the implication safe to assume in practice even as we note below a couple of cases in which the 0.90 quantile is (slightly) above 1.282, and acknowledge that subsequent research might reveal additional cases.

Let us elaborate. We have not formally proved that the 0.90 and 0.95 quantiles of Clark and McCracken's (2001) distribution obey (4.2). Rather, our observation is that the numerically computed quantiles obey (4.2). Also, while we have confidence in the code that computed the quantiles, we have not "proved" that the code used to generate the critical values is correct in any formal sense. Nor do we claim that sufficiently many simulations were done that there is near certainty that all the many digits in the tables are all correct. Indeed, so many simulations were done that with high probability some of the digits in some of the entries will be slightly off. Now, of the 400 sets of tabulated values, all 400 obey both inequalities in (4.2) for the recursive scheme, all 400 obey the upper inequality in (4.2) for the rolling scheme but "only" 396 of the 400 obey the lower inequality for the rolling scheme. The statement above that (4.2) holds "apart from a couple of exceptions" reflects the fact that in four cases the 0.90 quantile is 1.29, barely above the 1.282 value stated in the inequality.[6] Some other values are quite near 1.282, and it is possible that more extensive simulations intended to generate more accurate estimates of the quantiles would push some other values slightly above 1.282. It is our view that these or other possible corrections to the exact values in Clark and McCracken's (2001) table are very unlikely to undermine the practical relevance of interpreting a 1.282 critical value as defining a test of size somewhere between 0.05 and 0.10.

As well, it is possible that the critical values for values of P/R not tabulated strongly violate the inequalities. But while there is some minor wiggling up and down as one varies P/R across the 20 values stated above, there are no dramatic movements. So we consider it unlikely that critical values of P/R intermediate between tabulated ones will have markedly different critical values.

We therefore proceed on the understanding that use of a 1.282 critical value defines a test whose size is somewhere between 0.05 and 0.10, when the dimension of $Z_t \leqslant 20$ and for $P/R \leqslant 20.0$.

Recall that the 0.95 quantile for a normal distribution is 1.645. We note that inspection of the Clark and McCracken (2001) tables also reveals that apart from a handful of cases

$$0.95 \text{ quantile} \leqslant 1.645 \leqslant 0.99 \text{ quantile}. \tag{4.3}$$

[6]The values of P/R and the dimension of Z_t for these four cases happen to be (1) 5.0, 20; (2) 7.0, 18; (3) 7.0, 19; (4) 7.0, 20.

The upper inequality in (4.3) holds for all tabulated entries. The lower inequality is violated by 1 (recursive) or 14 (rolling) entries in which the 0.95 quantile is 1.65 or 1.66. Thus for one step ahead forecasts, tests using a critical value of 1.645 will define a test of size between 0.01 and 0.05 (approximately), for P and R sufficiently large.

While one-step ahead forecasts of conditionally homoskedastic errors are perhaps the leading example in practice, much finance data displays heteroskedasticity. And multistep predictions are common. Clark and McCracken (2005a) establish that when the dimension of Z_t is 1, the quantiles discussed above are still applicable even in the presence of conditional heteroskedasticity, and for multi-as well as one step ahead forecasts.

This leaves open inference when the dimension of Z_t is more than 1, and there are conditionally heteroskedastic and/or multistep forecasts. For the rolling scheme, it follows from Giacomini and White (2004) that if R is held fixed, and $P \to \infty, \hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1})$ obeys the usual law of large numbers and central limit theorem as $P \to \infty$:

$$
\begin{aligned}
&-2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) \to_p -2E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}),\\
&-2P^{-1/2}\left[\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) - E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})\right] \sim_A N(0,V),\\
&V = 4 \times \text{long run variance of } \hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}).
\end{aligned} \tag{4.4}
$$

The long run variance figures into V even for one step ahead forecast errors.[7]

In general, $E\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1})$ is non-zero. (An exception is the Clark and West (2005) environment in which $\beta_1^* \equiv 0$ and so $\hat{\beta}_{1t} \equiv 0$.) So under the null given in (3.3), as well as under the alternative given in (3.5), MSPE-adjusted will converge in probability to a nonzero value as $P \to \infty$ with R fixed. In light of the asymptotic result (4.4), there is, however, a straightforward interpretation of the usual t-statistic, in terms of confidence interval coverage. A p-value of (say) 0.15 means that an 85 percent confidence interval around the estimate of $E\hat{e}_{1t+1}(\hat{y}_{1t+1} - \hat{y}_{2t+1})$ contains zero. Suppose that our simulations cause us to report that (say) 18.4 percent of our t-statistics were above 1.282. Then had we constructed 90 percent confidence intervals, 81.6 percent of them would include zero.

The approximation that we have just discussed, which holds R fixed as P goes to infinity, thereby implying R/P goes to 0, may not be obviously appealing. Nonetheless, the R fixed approximation rationalizes the behavior of MSPE-adjusted (approximately normal) and MSPE-normal (not normal) for large but empirically relevant values of P/R (say, P/R of 2 or above).[8]

For MSPE-adjusted, how about if one considers the recursive scheme, for multistep forecasts and/or forecasts that are conditionally heteroskedastic and the dimension of Z_t is

[7] Giacomini and White (2004) propose what they call an unconditional test of the equality of the sample MSPE difference. They similarly state that the long run variance must be computed even for one step ahead forecasts. Their analysis departs from ours in that they maintain as a primitive assumption that the sample MSPE difference is centered at zero, while our null implies that the difference is shifted downwards, see the discussion below (3.8).

[8] As indicated in footnote 4, an alternative asymptotic approximation in which P/R goes to 0 is also not obviously appealing. Our simulation evidence finds that the R fixed approximation works better than the $P/R \to 0$ approximation, in the following sense: the R fixed approximation rationalizes the behavior of MSPE-adjusted (approximately normal) and MSPE-normal (not normal) for large but empirically relevant values of P/R (say, $P/R \geqslant 2$); the $P/R \to 0$ approximation rationalizes the behavior of MSPE-normal (theoretically approximately normal) only for small and empirically uncommon values of P/R (say, $P/R \leqslant 0.10$).

greater than 1? Here we return to the $R \to \infty$ and $P \to \infty$ asymptotics of Clark and McCracken (2001, 2005a). As stated above, the limiting distribution depends on data-specific parameters. So Clark and McCracken (2005a) propose constructing critical values via simulations of the asymptotic distribution, with certain parameters of the distribution chosen to match certain moments of the actual data. Our simulations also consider this statistic, which we call "*MSPE-adjusted, simulation critical values*" This is abbreviated in our tables as "*MSPE-adj. simul. cvs*".[9]

5. Test statistics

Let $\hat{e}_{1t+1}$ and $\hat{e}_{2t+1}$ be the one-step ahead forecast errors. Let $\hat{f}_{t+1} = \hat{e}_{1t+1}^2 - \hat{e}_{2t+1}^2$ for MSPE-normal, $\hat{f}_{t+1} = \hat{e}_{1t+1}^2 - \left[\hat{e}_{2t+1}^2 - (\hat{y}_{1t+1} - \hat{y}_{2t+1})^2\right]$ for MSPE-adjusted. Let $\bar{f}$ be the corresponding sample average, $\bar{f} = P^{-1}\Sigma_{t=R}^{T}\hat{f}_{t+1}$. Our test statistic is

$$\sqrt{P}\bar{f}/[\text{sample variance of } \hat{f}_{t+1} - \bar{f}]^{1/2}. \tag{5.1}$$

We also report simulation results on the out of sample test for nested models proposed by Chao et al. (2001) and generalized in Corradi and Swanson (2002). In the notation of (3.1) and (3.2), the null and sample moment used to test the null are

$$Ee_{1t}Z_t' = 0, \tag{5.2}$$

$$P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}Z_{t+1}' \quad \text{(CCS)}. \tag{5.3}$$

The χ^2-test statistic associated with (5.3) was adjusted for uncertainty due to estimation of regression parameters as described in Chao et al. (2001).

Bootstrap p-values were computed from a model-based, wild bootstrap, percentile-t method. In generating the bootstrap samples, we assumed correct specification of both equations of the two equation DGPs described in the next section, thus likely overstating the accuracy of the bootstrap in practice. For each simulation sample, 999 bootstrap samples were generated. The procedure was also used in Clark and West (2005), and details may be found in Section 4 of that paper. In the tables, we report bootstrap results under the label "*MSPE-adj.*: *bootstrap*". The appendix reports bootstrap results for MSPE-normal and CCS as well.

6. Simulation evidence

We use Monte Carlo simulations of simple bivariate data-generating processes to evaluate finite-sample size and power. We use two baseline DGPs, both of which incorporate features common in applications in which forecasts from estimated nested models are compared. In one DGP, which is motivated by asset-pricing applications, the variance of the predictand y_t is very high relative to the variance of the alternative model's additional predictors Z_t, and those additional predictors are highly persistent. In the second baseline DGP, which is motivated by macro applications, the parsimonious models's regression vector X_{1t} includes lags of the predictand y_t; the alternative model's Z_t

[9] What we call "MSPE-adjusted, simulations cvs" is called "Enc-t" in Clark and McCracken (2001, 2005a).

contains lags of an additional, persistent variable. We compare the tests listed in the previous section, for both the rolling and recursive estimation schemes.

6.1. Experimental design

The first DGP, meant to reflect asset pricing applications, takes a basic form widely used in studies of the properties of predictive regressions (see, for example Nelson and Kim, 1993; Stambaugh, 1999; Campbell, 2001; Tauchen, 2001):

$$\begin{aligned} &y_t = 0.5 + \gamma^* z_{t-1} + e_{1t}, X_{1t} = 1, X_{2t} = (1, z_{t-1})', z_t = 0.15 + 0.95 z_{t-1} + v_t, \\ &E_{t-1} e_{1t} = 0, E_{t-1} v_t = 0, \mathrm{var}(e_{1t}) = 18.0, \mathrm{var}(v_t) = 0.025, \mathrm{corr}(e_{1t}, v_t) = -0.75; \\ &\gamma^* = 0 \text{ in size experiments, } \gamma^* = 0.35 \text{ in power experiments.} \end{aligned} \tag{6.1}$$

DGP 1 is calibrated roughly to monthly excess returns in the S&P500 (y_t) and the dividend price ratio (z_t). While we focus on results for data generated from homoskedastic draws from the normal distribution, we extend DGP 1 to consider data with conditional heteroskedasticity–a feature often thought to characterize financial data. Select size results are reported for experiments in which e_t follows a GARCH(1,1) process, parameterized according to estimates for excess returns in the S&P500:

$$e_{1t} = \sqrt{h_t}\varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } N(0, 18), \quad h_t = 0.05 + 0.85 h_{t-1} + 0.1(e_{1t-1}^2/18). \tag{6.2}$$

Select results are also reported for experiments in which there is conditional heteroskedasticity in e_t, of a multiplicative form:

$$e_{1t} = \sqrt{h_t}\varepsilon_t, \varepsilon_t \sim \text{i.i.d. } N(0, 18), h_t = (z_{t-1} - Ez_t)^2/\sigma_z^2. \tag{6.3}$$

Note that both of these heteroskedasticity designs are parameterized so as to keep the unconditional mean and variance of y_t the same as in the homoskedastic case.

The second DGP is motivated by recent work on the predictive content of factor indexes of economic activity for output growth (examples include Stock and Watson, 2002, 2004; Marcellino et al., 2003; Shintani, 2005). The DGP is based on models estimated with quarterly data for 1967–2004 on GDP growth and the Federal Reserve Bank of Chicago's factor index of economic activity. For DGP 2, y_t corresponds to growth in GDP, and z_t corresponds to the Chicago Fed's factor index. The data generating process takes the following form:

$$\begin{aligned} &y_t = 2.237 + 0.261 y_{t-1} + \gamma_1^* z_{t-1} + \gamma_2^* z_{t-2} + \gamma_3^* z_{t-3} + \gamma_4^* z_{t-4} + e_{1t}, \\ &z_t = 0.804 z_{t-1} - 0.221 z_{t-2} + 0.226 z_{t-3} - 0.205 z_{t-4} + v_t, \\ &\mathrm{var}(e_{1t}) = 10.505, \quad \mathrm{var}(v_t) = 0.366, \quad \mathrm{cov}(e_{1t}, v_t) = 1.036, \\ &\gamma_i^* = 0, \quad i = 1, \ldots, 4, \text{ in size experiments;} \\ &\gamma_1^* = 3.363, \quad \gamma_2^* = -0.633, \quad \gamma_3^* = -0.377, \quad \gamma_4^* = -0.529 \text{ in power experiments.} \end{aligned} \tag{6.4}$$

To match the variety of settings that appear in empirical work, we consider a range of R and P values, with P both large and small relative to R. For the pseudo-macro DGP 2, we have in mind quarterly data, and consider $R = 80$, 120 and $P = 40$, 80, 120, 160. The comparable values for the pseudo-asset pricing DGP 1 are $R = 120$, 240 and $P = 120$, 240, 360, 720. For the given setting of R, a total of $R+160$ (or $R+720$ in our analysis of

"monthly" data) are generated. The initial observations on y and z are generated by a draw from a normal distribution whose variance–covariance matrix matches the unconditional variance–covariance matrix implied by the DGP. One-step ahead predictions are formed for observations $t = R+1$ through $R+160$ (or $R+720$), using models estimated with observations $t-R$ through $t-1$ (rolling) or observations 1 through $t-1$ (recursive). For each value of P, one-step ahead predictions are evaluated from $R+1$ through $R+P$. For multistep predictions of horizon τ, predictions are evaluated from $R+\tau$ through $R+P$, with the total number of predictions being $P-\tau+1$. The number of simulations is 5000.

For MSPE-normal and MSPE-adjusted "rejection" is defined as: the t-statistic is greater than $+1.282$. For CCS, we refer to the 0.90 quantiles of a $\chi^2(1)$ (DGP 1) or $\chi^2(4)$ (DGP 2) distribution. For MSPE-adj. simul. cvs., and MSPE-bootstrap, we define rejection as: the t-statistic is above the 0.90 quantile in the simulated or bootstrap distribution. An Appendix available on request from the authors contains results when we use a standard 0.05 cutoff (e.g., t-statistic cutoff of $+1.645$). We summarize below some results from that Appendix.

6.2. Simulation results

As discussed above, for MSPE-adjusted, our rejection rule defines a test of size between 0.05 and 0.10, where the size depends on the sampling scheme, dimension of Z_t and P/R.

Tables 1 and 2 presents results for homoskedastic (Table 1) and conditionally heteroskedastic (Table 2) data. The results for MSPE-adjusted are in good conformity with the asymptotic analysis presented above. Most notably, actual sizes fall between 0.05 and 0.10 in all 48 entries in the two tables. As well, sizes tend to be relatively close to 0.10 in ways that are consistent with the results in Clark and McCracken (2001, 2005a).[10] Specifically, sizes are closer to 0.10 than to 0.05 for rolling rather than recursive and for larger rather than smaller dimension of Z_t (DGP 2 rather than DGP 1).

As in Clark and McCracken (2001, 2005a), Clark and West (2005) and Corradi and Swanson (2005), MSPE-normal is seriously undersized. The median size is 0.008 across the 48 entries. Performance degrades (becomes more undersized) for larger P and for smaller R. This reflects the fact that MSPE normal has a negative mean and median. Recall that the numerator of the MSPE normal statistic is the difference in MSPEs, $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$, while the numerator in the MSPE adjusted statistic is $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.}) \equiv \hat{\sigma}_1^2 - \left(\hat{\sigma}_2^2 - P^{-1}\Sigma_{t=R}^{T}[\hat{y}_{1t+1} - \hat{y}_{2t+1}]^2\right)$ (see (3.9)). To illustrate the mean and median bias in MSPE normal, consider DGP 1, $R = 120$ and $P = 720$ (Table 1, panel 1A). Across 5000 simulations, the mean and median value of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is -0.24, while the mean and median values of $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - \text{adj.})$ are 0.01 and 0.02 (not reported in the table). (To scale these figures, it may be helpful to recall that the population MSPE is 18.0.) Across simulations, the implied mean value of the squared difference in fitted values $P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$ is $0.25 (= 0.01 - (-0.24))$.

Thus, the behavior of MSPE-normal is consistent with the test statistic being dominated by the squared differences in fitted values (the term $-P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$ on the r.h.s. of (3.8)). Since this term is negative, and since we are using one-tailed tests that only reject when the test statistic is sufficiently positive, the test is undersized. Given R, the

[10] The occasional oversizing Clark and McCracken (2001, 2005a) find arises when data-determined lag selection yields significantly misspecified null forecasting models.

Table 1
Empirical size: 1-step ahead forecasts, nominal size = 10%

A. Rolling regressions

	1. DGP 1, $R = 120$				2. DGP 1, $R = 240$			
	$P = 120$	$P = 240$	$P = 360$	$P = 720$	$P = 120$	$P = 240$	$P = 360$	$P = 720$
MSPE-adjusted	0.085	0.078	0.080	0.091	0.079	0.061	0.062	0.065
MSPE-normal	0.018	0.006	0.001	0.000	0.035	0.012	0.007	0.001
CCS	0.106	0.109	0.109	0.132	0.106	0.109	0.100	0.106
MSPE-adj.:simul. cvs	0.137	0.129	0.129	0.125	0.106	0.110	0.109	0.111
MSPE-adj: bootstrap	0.102	0.103	0.101	0.106	0.092	0.090	0.090	0.094
	3. DGP 2, $R = 80$				4. DGP 2, $R = 120$			
	$P = 40$	$P = 80$	$P = 120$	$P = 160$	$P = 40$	$P = 80$	$P = 120$	$P = 160$
MSPE-adjusted	0.094	0.090	0.085	0.084	0.093	0.084	0.080	0.078
MSPE-normal	0.018	0.002	0.000	0.000	0.028	0.008	0.005	0.001
CCS	0.146	0.117	0.119	0.111	0.149	0.120	0.114	0.104
MSPE-adj.:simul. cvs	0.119	0.115	0.117	0.111	0.116	0.112	0.106	0.103
MSPE-adj: bootstrap	0.094	0.098	0.099	0.096	0.091	0.093	0.093	0.093

B. Recursive regressions

	1. DGP 1, $R = 120$				2. DGP 1, $R = 240$			
	$P = 120$	$P = 240$	$P = 360$	$P = 720$	$P = 120$	$P = 240$	$P = 360$	$P = 720$
MSPE-adjusted	0.085	0.070	0.062	0.054	0.075	0.063	0.058	0.055
MSPE-normal	0.028	0.012	0.009	0.003	0.037	0.022	0.014	0.008
CCS	0.107	0.108	0.107	0.101	0.107	0.103	0.105	0.099
MSPE-adj.:simul. cvs	0.134	0.120	0.107	0.106	0.112	0.111	0.104	0.101
MSPE-adj: bootstrap	0.104	0.100	0.098	0.097	0.092	0.094	0.092	0.098
	3. DGP 2, $R = 80$				4. DGP 2, $R = 120$			
	$P = 40$	$P = 80$	$P = 120$	$P = 160$	$P = 40$	$P = 80$	$P = 120$	$P = 160$
MSPE-adjusted	0.088	0.087	0.081	0.076	0.090	0.086	0.082	0.075
MSPE-normal	0.024	0.010	0.006	0.002	0.030	0.013	0.008	0.006
CCS	0.146	0.119	0.112	0.103	0.149	0.116	0.111	0.105
MSPE-adj.:simul. cvs	0.121	0.114	0.110	0.109	0.112	0.111	0.113	0.098
MSPE-adj: bootstrap	0.087	0.096	0.093	0.093	0.088	0.093	0.096	0.089

Notes: (1) In DGP 1, the predictand y_{t+1} is i.i.d. normal around a nonzero mean; the alternative model's predictor z_t follows an AR(1) with parameter 0.95. In DGP 2, y_{t+1} follows an AR(1) with parameters given in (6.4); the alternative model includes lags of an AR(4) variable z_t along with the lag of y_t, again with parameters given in (6.4). In each simulation, and for each DGP, one step ahead forecasts of y_{t+1} are formed from each of the two models, using least squares regressions. (2) R is the size of the rolling regression sample (panel A), or the smallest regression sample (panel B). P is the number of out-of-sample predictions. (3) *MSPE–normal* is the difference in mean squared prediction errors, see (3.7); *MSPE–adjusted* adjusts the difference in mean squared prediction errors to account for the additional predictors in the alternative models, see (3.9); *CCS* is the Chao et al. (2001) statistic testing whether model 1 forecasts are uncorrelated with the additional predictors in model 2, see (5.3); *MSPE-adj.*: *simul. cvs* uses simulations of the non-standard limiting distribution in Clark and McCracken (2005a) to compute critical values for the MSPE-adjusted statistic; *MSPE-adj.*: *bootstrap* uses a percentile-t wild bootstrap, with 999 replications per simulation sample. (4) The number of simulations is 5000. For MSPE-adjusted and MSPE-normal, the table reports the fraction of simulations in which each test statistic was greater than 1.282, which is the standard normal critical value for a one-sided test at the 10% level. For example, panel A1, $P = 120$, MSPE-adjusted, 425 test statistics were greater than 1.282. This led to the figure of 0.085 given in the table. For CCS, sizes were computed using $\chi^2(1)$ (DGP 1) or $\chi^2(4)$ (DGP 2) critical values. (5) For large P and R: MSPE-adjusted has size between 0.05 and 0.10, MSPE-normal has size below 0.10; the other three statistics have size 0.10.

Table 2
Empirical size: DGP 1 with heteroskedasticity, nominal size = 10%, 1-step ahead forecasts, $R = 120$

	1. GARCH				2. Multiplicative			
	$P = 120$	$P = 240$	$P = 360$	$P = 720$	$P = 120$	$P = 240$	$P = 360$	$P = 720$
A. Rolling regressions								
MSPE-adjusted	0.080	0.073	0.078	0.086	0.111	0.094	0.086	0.083
MSPE-normal	0.017	0.004	0.001	0.000	0.015	0.004	0.001	0.000
CCS	0.106	0.102	0.107	0.122	0.085	0.085	0.091	0.095
MSPE-adj.:simul. cvs	0.135	0.123	0.124	0.120	0.174	0.153	0.134	0.115
MSPE-adj: bootstrap	0.099	0.093	0.099	0.101	0.109	0.108	0.100	0.095
B. Recursive regressions								
MSPE-adjusted	0.076	0.064	0.063	0.057	0.107	0.086	0.085	0.070
MSPE-normal	0.028	0.011	0.009	0.004	0.030	0.017	0.010	0.005
CCS	0.107	0.100	0.101	0.094	0.088	0.086	0.087	0.091
MSPE-adj.:simul. cvs	0.129	0.117	0.109	0.105	0.167	0.155	0.138	0.127
MSPE-adj: bootstrap	0.096	0.094	0.099	0.093	0.115	0.111	0.115	0.112

Notes: (1) See the notes to Table 1. (2) Panel A, the predictand y_{t+1} is a GARCH process, with the parameterization given in Eq. (6.2). In panel B, the predictand y_{t+1} has conditional heteroskedasticity of the form given in Eq. (6.3), in which the conditional variance at t is a function of z_{t-1}^2.

expectation of $(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$ is fixed, say $\hat{y}(R)$. If we hold R fixed, as in asymptotics proposed by Giacomini and White (2004), then as P gets bigger a law of large numbers makes $-P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$ collapse on $-\hat{y}(R)$. This makes the probability of a negative test statistic larger and larger. As R gets bigger (given P) $\hat{y}(R)$ moves towards zero (since as $R \to \infty$, $\hat{y}_{1t+1} - \hat{y}_{2t+1} \to_p 0$), thus explaining the improved size with bigger R.

Fig. 1 presents smoothed density estimates of the MSPE-normal and MSPE-adjusted test statistics for DGP 1, $R = 240$, $P = 120$, 240, and 720. ($P = 360$ was omitted to keep the plots legible.) Figs. 1A and B present results for the rolling scheme, depicting densities associated with the results presented in the first two lines of panel A2 in Table 1. Figs. 1C and D do the same for the recursive scheme; the associated results are presented in the first two lines of panel B2 in Table 1.

That MSPE-adjusted and MSPE are undersized is clear in all four panels: our one-tailed tests, which reject only if the t-statistic is greater than 1.282, will clearly reject less than 10 percent of the time given the leftward shift in the distributions. It is equally clear, however, that MSPE-adjusted will be far better sized than MSPE-normal, because of the sharper leftward shift in MSPE-normal. Figs. 1A and C illustrate how the distribution of MSPE-normal piles up on what we called $\hat{y}(R)$ as P increases.

Since we have argued that there is no good reason to use asymptotic normal critical values with MSPE-normal, it is perhaps no surprise that MSPE-adjusted does much better than MSPE-normal. But the performance of MSPE-adjusted, while not matching up to the ideal standard of empirical sizes of exactly 0.10, does credibly against other competitors. We see in Tables 1 and 2 that the CCS statistic is very nicely sized in DGP 1, but a bit oversized in DGP 2 (panels A3, A4, B3, B4 in Table 1). MSPE with simulation-based critical values is slightly oversized in all DGPs. MSPE with bootstrap p-values does very well.

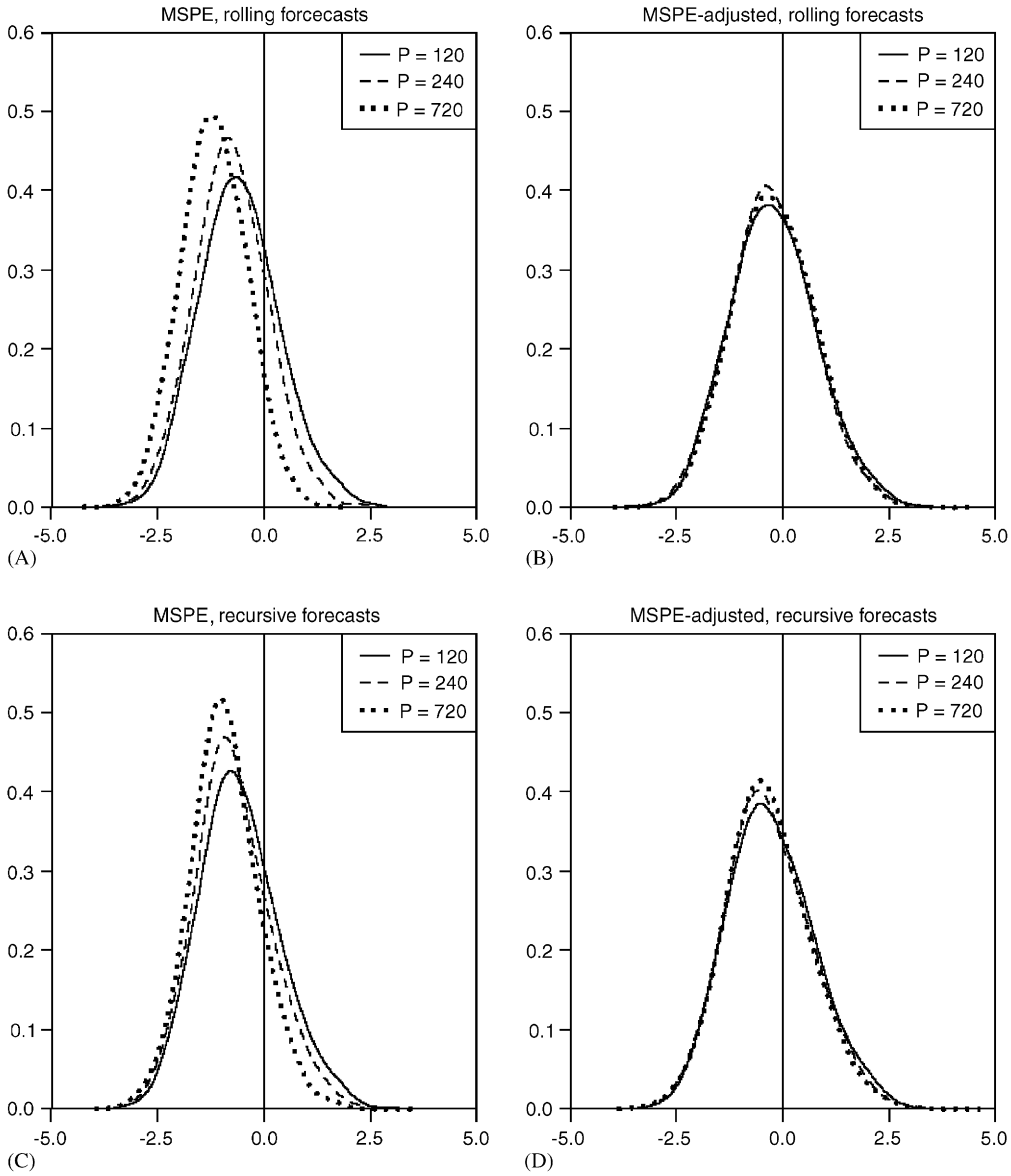

Fig. 1. Null densities of simulated tests, DGP 1, $R = 240$, P varying.

Perhaps a good summary statistic to compare the five test statistics is the median empirical size. Across all 48 entries, median empirical sizes were: MPSE-adjusted: .080; MSPE-normal: 0.008; CCS: 0.107; MSPE-adj. simul. cvs: 0.115; MSPE-adj. bootstrap: 0.096.

Results for tests using a critical value of $+1.645$ are presented in the not for publication Appendix. They tell the same story. In accordance with the asymptotic theory, for MSPE-adjusted, the 48 sets of simulations generally (with three exceptions) yielded sizes between 0.01 and 0.05 (the actual range was from 0.027 to 0.059). The median size was 0.041.

Table 3
Size-adjusted power: 1-step ahead forecasts, size = 10%

A. Rolling regressions

	1. DGP 1, $R = 120$				2. DGP 1, $R = 240$			
	$P = 120$	$P = 240$	$P = 360$	$P = 720$	$P = 120$	$P = 240$	$P = 360$	$P = 720$
MSPE-adjusted	0.162	0.194	0.218	0.257	0.183	0.221	0.242	0.303
MSPE-normal	0.153	0.180	0.193	0.233	0.166	0.196	0.208	0.269
CCS	0.064	0.074	0.087	0.158	0.058	0.055	0.060	0.091
	3. DGP 2, $R = 80$				4. DGP 2, $R = 120$			
	$P = 40$	$P = 80$	$P = 120$	$P = 160$	$P = 40$	$P = 80$	$P = 120$	$P = 160$
MSPE-adjusted	0.964	0.999	1.000	1.000	0.972	1.000	1.000	1.000
MSPE-normal	0.845	0.985	0.999	1.000	0.828	0.981	0.998	1.000
CCS	0.608	0.941	0.997	1.000	0.591	0.934	0.995	1.000

B. Recursive regressions

	1. DGP 1, $R = 120$				2. DGP 1, $R = 240$			
	$P = 120$	$P = 240$	$P = 360$	$P = 720$	$P = 120$	$P = 240$	$P = 360$	$P = 720$
MSPE-adjusted	0.172	0.228	0.268	0.354	0.191	0.231	0.282	0.381
MSPE-normal	0.169	0.202	0.234	0.328	0.182	0.208	0.250	0.336
CCS	0.060	0.055	0.058	0.059	0.052	0.054	0.048	0.056
	3. DGP 2, $R = 80$				4. DGP 2, $R = 120$			
	$P = 40$	$P = 80$	$P = 120$	$P = 160$	$P = 40$	$P = 80$	$P = 120$	$P = 160$
MSPE-adjusted	0.972	1.000	1.000	1.000	0.973	1.000	1.000	1.000
MSPE-normal	0.842	0.980	0.998	1.000	0.827	0.975	0.997	1.000
CCS	0.606	0.927	0.992	1.000	0.593	0.927	0.991	1.000

Notes: (1) In panels A1, A2, B1 and B2, the DGP is defined in equation 6.1, with the nonzero value of γ^* given in that equation. In panels A3, A4, B3 and B4, the DGP is defined in (6.4), with nonzero values of γ_i^* given in (6.4). (2) Power is calculated by comparing the test statistics against simulation critical values, calculated as the 90th percentile of the distributions of the statistics in the corresponding size experiment reported in Table 1. Because "MSPE-adjusted," "MSPE-adj. simul. cvs" and "MSPE-adj: bootstrap" use the same test statistic, size adjusted power is identical for the three.

Median sizes for other test statistics were: MSPE-normal: 0.003; CCS: 0.055; MSPE-adj. simul. cvs.: 0.061; MSPE-adj. bootstrap: 0.048.

Table 3 presents results on size-adjusted power, for one step ahead forecasts, and for the conditionally homoskedastic data generating processes also used in Table 1. As explained in the notes to the tables, the entry "MSPE-adjusted" applies to the "MSPE-adjusted," "MSPE-adj. simul. cvs" and "MSPE-adj bootstrap" entries in Table 1 because size adjusted power is identical for the three.

In DGP 1, size adjusted power is best for MSPE-adjusted, worst for CCS, with MSPE-normal in the middle. In DGP 2, power is best for MSPE-adjusted, worst for CCS, with MSPE-normal falling in the middle.

In practice, unadjusted power may be more relevant than size adjusted power. The size adjustment involves computing critical values by Monte Carlo methods. If a researcher completed such an exercise, the researcher would likely use the simulation rather than asymptotic critical values.

Unadjusted power is reported in detail in the appendix. On balance, the ranking from best to worst power is: MSPE-adj. simul. cvs, MSPE-adj. bootstrap, MSPE-adjusted, CCS, MSPE-normal. The differences between the first three are small. The difference between MSPE-normal and the other tests is huge. To illustrate, consider DGP 1, $R = 120$, $P = 360$. Unadjusted power is

MSPE adjusted	MSPE-normal	CCS	MSPE-adj. simul.cvs	MSPE-adj. bootstrap
0.190	0.031	0.061	0.280	0.270

(6.5)

For DGP 1, even the best unadjusted power is not good. This essentially reflects the fact that there is not much predictability in asset prices. In our calibration, the MSPE of the alternative model is about 5 percent lower than that of the null model (i.e., the R^2 in the alternative model is about 0.05). With such a small amount of predictability, it will take many, many observations to have high probability of rejecting the null. (By contrast, unadjusted power for DGP 2 was above 0.8 for most choices of P and R.)

In summary, of the three statistics that do not require simulations to compute critical values (MSPE-normal, CCS, and MSPE-adjusted), MSPE-normal has the worst power and size, while MSPE-adjusted cannot be beaten in terms of either size or power.

If we turn to statistics that do involve simulations, MSPE-adjusted is no longer undominated in terms of size and power. A bootstrap yields improvements in size relative to use of asymptotic normal critical values for MSPE-adjusted (median size of 0.096 rather than 0.080 for nominal 0.10 tests, median size of 0.048 rather than 0.041 for nominal 0.05 tests). While we do not belittle such improvements, we repeat that our bootstrap assumed knowledge of the correct specification (though not parameters) of the processes for y_t and z_t and thus our bootstrap results may be a bit generous. And, whether or not the results are a bit generous to the bootstrap, we observe that our simulations and asymptotic analysis indicate that for MSPE-adjusted, nominal 0.10 tests can be relied upon to deliver tests of actual size between 0.05 and 0.10, nominal 0.05 tests can be relied upon to deliver tests of actual size between 0.01 and 0.05. As well, power is roughly comparable when one uses bootstrapped ("MSPE-adj. bootstrap") or asymptotic normal ("MSPE-adjusted") critical values. Thus, the computational simplicity of MSPE-adjusted may make this statistic appealing in many applications as a simple way to deliver an approximately normal test statistic.

7. Empirical example

To illustrate our approach, we apply the MSPE-adjusted, MSPE-normal, and CCS tests to one-month ahead forecasts of excess stock returns and one quarter ahead forecasts of GDP growth. In the stock return application, the null model posits that the excess return on the S&P 500 is unpredictable around a time invariant mean. The alternative model, widely used in studies of the predictability of stock returns, relates the excess return to a constant and the dividend-price ratio. We calculated the excess return and dividend-price

ratio following the conventions of Pesaran and Timmermann (1995), using end-of-month stock prices taken from the Federal Reserve Board of Governors' FAME database, monthly dividends from Global Insight's S&P databank, and the one-month Fama/French interest rate series from Kenneth French's website. The initial sample runs from January 1954 through December 1963, so $R = 120$ months. Predictions run from January 1964 through December 2004, so the number of predictions is $P = 492$. Although not reported in the interest of brevity, full sample estimates of our excess return models are comparable to those reported in the literature: a (weakly) significantly negative coefficient on the dividend-price ratio and a small adjusted R^2.

In the GDP growth application, the null model is an AR(1) (including a constant). The alternative model, drawn from recent studies of the predictive content of factor indexes of the business cycle cited in the previous section, relates US GDP growth to a constant, one lag of GDP growth, and four lags of the Chicago Fed's national activity index. The GDP data were obtained from the Board of Governors' FAME database; the factor index (a quarterly average of the underlying monthly series) was taken from the Chicago Fed's

Table 4
Forecasts of monthly excess stock returns and quarterly GDP growth

(1) Predictand	(2) Prediction sample	(3) $\hat{\sigma}_1^2$	(4) $\hat{\sigma}_2^2$	(5) adj.	(6) $\hat{\sigma}_2^2$ – adj.	(7) MSPE-normal	(8) MSPE-adj.	(9) CCS
A. Rolling regressions								
(1) Excess stock return	Jan. 1964–Dec. 2004	18.92	19.57	0.67	18.90	−0.66 (0.33) *−2.00*	0.01 (0.32) *0.04*	1.28
(2) GDP growth	1985:Q1–2004:Q4	3.89	3.93	1.09	2.84	−0.04 (0.49) *−0.09**	1.04 (0.50) *2.07***	4.72
B. Recursive regressions								
(1) Excess stock return	Jan. 1964–Dec. 2004	18.91	19.14	0.68	18.46	−0.23 (0.38) *−0.63*	0.45 (0.38) *1.17**	0.14
(2)GDP growth	1985:Q1–2004:Q4	3.80	3.67	1.01	2.66	0.12 (0.49) *0.25**	1.14 (0.54) *2.10***	4.84

Notes: (1) In column (3), $\hat{\sigma}_1^2$ is the out of sample MSPE of the parsimonious model. For excess stock returns (return on S and P 500, less one month bond yield), the parsimonious model posits returns to be unpredictable around a time invariant mean. For GDP growth, the parsimonious model is a univariate AR(1). (2) In column (4), $\hat{\sigma}_2^2$ is the out of sample MSPE of an alternative larger model. For stock returns, the larger model includes a lag of the dividend-price ratio. For GDP growth, the larger model includes four lags of the Federal Reserve Bank of Chicago's factor index. (3) All forecasts are one step ahead. The model estimation start dates are January 1954 (stock returns) and 1968:Q2 (GDP growth). R is 120 months (stock returns) or 67 quarters (GDP growth). The number of predictions P is 492 (stock returns) or 80 (GDP growth). (4) In column (5), "adj." is the adjustment term $P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$, where $\hat{y}_{1t+1} - \hat{y}_{2t+1}$ is the difference between forecasts of the two models. In column (6), "$\hat{\sigma}_2^2$ – adj." is the difference between column (4) and column (5). (5) For each predictand, column (7) presents a point estimate of the difference in MSPEs (i.e., the difference between columns (3) and (4)), an asymptotic standard error in parentheses, and a t-statistic in italics. Column (8) does the same, but relying on the difference between columns (3) and (6). Figures may not add, due to rounding. (6) Column (9) presents the $\chi^2(1)$ (stock return) or $\chi^2(4)$ (GDP growth) statistics for the Chao et al. statistic (5.3). (7) **denotes test statistics significant at the 5 percent level according to both standard normal and Clark and McCracken's (2005a) asymptotic critical values; *denotes a test statistic significant at the 10 percent level according to Clark and McCracken (2001, 2005a).

web site. The initial sample runs from 1968:Q2 through 1984:Q4, so $R = 67$ quarters. Predictions run from 1985:Q1 through 2004:Q4, so the number of predictions is $P = 80$. Full sample estimates of the competing forecasting models indicate the activity index has significant explanatory power for GDP growth (with higher index values predicting higher GDP growth).

Table 4 contains our results. The table reflects the common difficulty of beating, in MSPE, parsimonious null models. In the stock return application, the MSPE of the model with the dividend-price ratio ($\hat{\sigma}_2^2 = 19.57$ for rolling, 19.14 for recursive) is above the MSPE of the model with just a constant ($\hat{\sigma}_1^2 = 18.92$ for rolling, 18.91 for recursive), for both rolling and recursive regressions. In the GDP growth example, the MSPE of the model with the activity index ($\hat{\sigma}_2^2 = 3.93$ for rolling, 3.67 for recursive) is slightly above the MSPE of the AR(1) model ($\hat{\sigma}_1^2 = 3.89$) in the rolling regression, slightly below in the recursive regression ($\hat{\sigma}_1^2 = 3.80$). Accordingly, without even calculating standard errors, we know that with the possible exception of the GDP growth example, recursive, use of the simple MSPE test with standard normal critical values (''MSPE-normal'') with a one tailed test will fail to reject the null model. We see in Panel B2, column (7) that even for GDP growth, recursive, the MSPE-normal test also fails to reject.

We have given analytical and simulation evidence that MSPE-normal is seriously undersized. For the stock return data, rolling, using either asymptotic normal or critical values from the Clark and McCracken (2001) table on the web, we continue to fail to reject the null even after adjustment (t-statistic is 0.04). For recursive, the t-statistic of 1.17 is below the 1.282 normal critical value but above the 0.90 quantile tabulated by Clark and McCracken (2001). Hence there is some statistical evidence against the null of no stock return predictability. For GDP growth, though, the adjustment leads to t-statistics of about 2.1 for both rolling and recursive forecasts, allowing rejection at a significance level between 0.01 and 0.05 (see Eq. (4.3)). Reference to the relevant Clark and McCracken quantiles also indicates rejection at significance level between 0.01 and 0.05. As well, for the recursive scheme, comparing the MSPE-normal test against asymptotic critical values simulated with the method of Clark and McCracken (2005a) does lead to a (weak) rejection of the null AR(1) model.

The results for our adjusted MSPE test highlight the potential for noise associated with the additional parameters of the alternative model to create an upward shift in the model's MSPE large enough that the null model has a lower MSPE even when the alternative model is true. The estimated adjustments in column (5) of Table 7 correspond to the term $P^{-1}\Sigma_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$. The adjustment is 0.67 or 0.68 for stock return forecasts (corresponding to about 3–4 percent of the alternative model's MSPE) and 1.01–1.09 for GDP growth forecasts (or roughly 25 percent). In the case of stock returns, the adjustment gives the alternative model a small advantage over the null model, but the adjustment is not large enough to cause the null model to be rejected. For GDP growth, though, the adjustment is large enough to not only give the alternative model an advantage over the null model, but also to cause the null model to be soundly rejected: the MSPE-adjusted test rejects the null model when compared against both standard normal and Clark and McCracken (2005a) simulated critical values.

Thus, while the unadjusted MSPE test would seem to support the null models of stock returns and GDP growth, our MSPE-adjusted test, which adjusts for the additional parameter noise in the alternative model, provides some evidence—more so for GDP growth than stock returns—in favor of alternative models. That is, in rolling regressions

(panel A) the univariate autoregressive model for GDP growth has a lower MSPE than does the bivariate model that includes the factor index. Nonetheless, after accounting for estimation noise in the bivariate model, there is strong evidence that a factor index of economic activity has additional predictive content for growth. Such a result underscores the practical relevance of our MSPE-adjusted statistic in MSPE comparisons of nested models.

8. Conclusions

Forecast evaluation often compares the mean squared prediction error of a parsimonious null model that is nested in a larger, and less parsimonious, model. Under the null that the parsimonious null model generates the data, the larger model introduces noise into its forecasts by attempting to estimate parameters whose population values are zero. This implies that the mean squared prediction error from the parsimonious model is expected to be *smaller* than that of the larger model.

We describe how to adjust mean-squared errors to account for this noise, producing what we call *MSPE-adjusted*. We recommend then constructing the usual *t*-statistics and rejection regions to test whether the adjusted difference in mean squared errors is zero. We refer to the quantiles of the nonstandard distribution tabulated in Clark and McCracken (2001, 2005a) to argue that this will result in modestly undersized tests: one-sided tests using 1.282 as the critical value will, in large samples, have actual size somewhere between 0.05 and 0.10; one sided tests using 1.645 will have size between 0.01 and 0.05. Simulations support our recommended procedure.

Acknowledgements

West thanks the National Science Foundation for financial support. We thank Pablo M. Pincheira-Brown, Philip Hans Franses, Taisuke Nakata, Norm Swanson, participants in a session at the January 2006 meeting of the Econometric Society and two anonymous referees for helpful comments. 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.

References

Ang, A., Piazzesi, M., Wei, M., 2004. What does the yield curve tell us about GDP growth? Journal of Econometrics 131, 359–403.

Ashley, R., Granger, C.W.J., Schmalensee, R., 1980. Advertising and aggregate consumption: an analysis of causality. Econometrica 48, 1149–1168.

Campbell, J.Y., 2001. Why long horizons? A study of power against persistent alternatives. Journal of Empirical Finance 8, 459–491.

Campbell, J.Y., Thompson, S.B., 2005. Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average? Harvard University, Cambridge, MA (manuscript).

Chao, J., Corradi, V., Swanson, N.R., 2001. Out-of-sample tests for Granger Causality. Macroeconomic Dynamics 5, 598–620.

Chong, Y.Y., Hendry, D.F., 1986. Econometric evaluation of linear macro-economic models. Review of Economic Studies 53, 671–690.

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., 2005a. Evaluating direct multistep forecasts. Econometric Reviews 24, 369–404.
Clark, T.E., McCracken, M.W., 2005b. The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence. Journal of Money, Credit, and Banking forthcoming.
Clark, T.E., West, K.D., 2005. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Journal of Econometrics forthcoming.
Cooper, M.R.C., Gutierrez Jr., Marcum, W., 2005. On the predictability of stock returns in real time. Journal of Business 78, 469–500.
Corradi, V., Swanson, N.R., 2002. A consistent test for nonlinear predictive ability. Journal of Econometrics 110, 353–381.
Corradi, V., Swanson, N.R., 2005. Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. International Economic Review forthcoming.
Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. Journal of Econometrics 130, 337–364.
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.H., 2005. News and noise in G-7 GDP announcements. Journal of Money, Credit and Banking 37, 403–420.
Giacomini, R., White, H., 2004. Tests of conditional predictive ability. University of California, San Diego Econometrica, forthcoming.
Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Management Science 49, 639–654.
Harvey, D.I., Leybourne, S.J., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.
Kilian, L., 1999. Exchange rates and monetary fundamentals: what do we learn from long-horizon regressions? Journal of Applied Econometrics 14, 491–510.
Lettau, M., Ludvigson, S., 2001. Consumption, wealth, and expected stock returns. Journal of Finance 56, 815–849.
Marcellino, M., Stock, J.H., Watson, M.W., 2003. Macroeconomic forecasting in the Euro area: country-specific versus area-wide information. European Economic Review 47, 1–18.
Marcellino, M., Stock, J.H., Watson, M.W., 2004. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics forthcoming.
Mark, N., 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. American Economic Review 85, 201–218.
McCracken, M.W., 2004. Asymptotics for Out of Sample Tests of Causality. University of Missouri manuscript.
Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: the role of small sample bias. Journal of Finance 48, 641–661.
Orphanides, A., van Norden, S., 2005. The reliability of inflation forecasts based on output gap estimates in real time. Journal of Money, Credit and Banking 37, 583–601.
Rapach, D.E., Weber, C.E., 2004. Financial variables and the simulated out-of-sample forecastability of US output growth since 1985: an encompassing approach. Economic Inquiry 42, 717–738.
Shintani, M., 2005. Nonlinear forecasting analysis using diffusion indexes: an application to Japan. Journal of Money, Credit, and Banking 37, 517–538.
Stambaugh, R.F., 1999. Predictive regressions. Journal of Financial Economics 54, 375–421.
Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20, 147–162.
Stock, J.H., Watson, M.W., 2003. Forecasting output and inflation: the role of asset prices. Journal of Economic Literature 41, 788–829.
Stock, J.H., Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405–430.
Tauchen, G., 2001. The bias of tests for a risk premium in forward exchange rates. Journal of Empirical Finance 8, 695–704.
West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.
West, K.D., 2001. Tests of forecast encompassing when forecasts depend on estimated regression parameters. Journal of Business and Economic Statistics 19, 29–33.
West, K.D., 2005. Forecast Evaluation. University of Wisconsin (manuscript).
West, K.D., McCracken, M.W., 1998. Regression based tests of predictive ability. International Economic Review 39, 817–840.