APPROXIMATELY NORMAL TESTS FOR EQUAL PREDICTIVE ACCURACY IN NESTED MODELS

Todd E. Clark
Federal Reserve Bank of Kansas City

Kenneth D. West
University of Wisconsin

June 2005

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 (1996)) to test whether the adjusted mean squared error difference is zero. We refer to nonstandard limiting distributions derived in Clark and McCracken (2001, 2005a) 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.

West thanks the National Science Foundation for financial support. We thank Pablo M. Pincheira-Brown and Taisuke Nakata 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.
Note to referees: the Appendix that is referenced in the paper has been submitted along with the paper.

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). A closely related measure also in widespread use is correlation between the parsimonious model's prediction error and larger model's forecasts (encompassing). In this paper we explore the behavior of standard normal inference for MSPE and encompassing statistics 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*.

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 (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. The Clark and McCracken (2001, 2005a) asymptotics and simulation quantiles indicate that under certain circumstances, tests using the 10 percent normal critical value (1.282, for one-sided tests) will have actual size between 5 and 10 percent, while those using the 5 percent normal critical value (1.645, for one sided tests) will yield actual size between 1 and 5 percent. The circumstances under which this applies are: (1)one step ahead conditionally homoskedastic forecast errors, or (2)multistep and/or conditionally heteroskedastic forecast errors when the larger model relies on exactly one more parameter than the smaller model. The second condition may seem special, but many asset pricing applications in fact involve MSPE comparisons in which the larger model includes just one more parameter.[1] And even if these circumstances do not apply, simulations suggest that the normal approximation will work reasonably well. But, formally, comparing standard normal critical values against Clark and McCracken's quantiles requires the conditions just noted.

Our simulations show that these quantiles are applicable with samples of size typically available. We complete 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 these simulations, use of the .10 normal critical value of 1.282 resulted in 44 tests with actual size between .05 and .10, and 4 tests with size slightly larger than .10. The four tests with size that fall above the .10 upper bound predicted by our theory all involve relatively small sample sizes. The median size across the 48 sets is 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 .05 normal critical value of 1.645: 44 tests have actual size between .01 and .05, while 4 with small sample sizes were slightly oversized. The median size is about .04. These results are consistent with the simulations in Clark and McCracken (2001, 2005a).

We focus on MSPE because empirical practice of applied researchers indicates that MSPE or root MSPE are objects of great interest.[2] All simulations also looked at standard normal inference for the raw (unadjusted) difference in MSPEs. We call this "*MSPE-normal*" in our tables. Consistent with the asymptotic theory and simulations in McCracken (2004) and Clark and McCracken (2001, 2005a), MSPE-normal performed abysmally. For the one-step ahead forecasts and nominal .10 tests, the median size across 48 sets of simulations was less than 0.01, for example.

The widespread use of MSPE might be ill advised if related moments lead easily to reliable discrimination between nested models. Our simulations therefore also examine an encompassing statistic proposed by Chong and Hendry (1986) ("*CH*",in our tables) and a certain statistic for nested models proposed by Chao, Corradi and Swanson (2001) ("*CCS*", in our tables). Asymptotic normality of CH follows from conditions and arguments such as those in West (1996); see Chao et al. (2001) for conditions under which CCS has a standard limiting distribution. We find that CH performs more poorly than MSPE-adjusted. In almost all of the 48 simulations, this statistic is more undersized than MSPE-adjusted; the median size is about 0.06. CCS performs a little better than does MSPE-adjusted. Its median size is 0.11. In terms of size adjusted power, MSPE-adjusted and CH perform better than CCS and MSPE-normal. For raw (unadjusted) power, MSPE-adjusted performs better than any of the other

statistics.

We also briefly consider tests relating to multistep forecasts. CH and CCS are asymptotically normal or chi-squared under suitable conditions. As noted above, the quantiles from Clark and McCracken (2001, 2005a) can be used to rationalize use of standard normal critical values for MSPE-adjusted, and multistep forecasts, when the larger model relies on exactly one more parameter than the smaller model. We apply standard normal inference to MSPE-adjusted using one DGP that conforms to the requirement of a single extra parameter in the larger model, and one that does not. There is little apparent difference in performance of MSPE-adjusted across the two DGPs. On balance, MSPE-adjusted performs a little better than CH and CCS, a lot better than MSPE-normal. The performance of CH and MSPE-normal improves, that of CCS and MSPE-adjusted degrades.

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). This prior work has shown a bootstrap to be reliable (at least with models reasonably close to being correctly specified)–reliable enough that, in the interest of brevity, we omit bootstrap results from this paper. Another method, which we do include in this paper, 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 results in modest improvements in size relative to MSPE-adjusted (median size across 48 sets of simulations = 0.11).

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 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 the submitted 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 mean squared prediction error 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, $\hat{\sigma}_1^2=P^{-1}\sum(y_{t+\tau}-\hat{y}_{1t,t+\tau})^2$, $\hat{\sigma}_2^2=P^{-1}\sum(y_{t+\tau}-\hat{y}_{2t,t+\tau})^2$, $\hat{\sigma}_2^2$-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 $\hat{\sigma}_1^2-(\hat{\sigma}_2^2$-adj.), rejecting if this difference is sufficiently positive. Note that $(\hat{\sigma}_2^2$-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

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

Now, $\hat{\sigma}_1^2-(\hat{\sigma}_2^2$-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 .10 test) or +1.645 (for a one sided .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}:

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

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

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 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,

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

The commonly examined implications of (3.3) include:

(3.4) $Ee_{1t}^2-Ee_{2t}^2=0$, (equal MSPE)

(3.5) $Ee_{1t}(X_{2t}'\beta_2^*)=0$, (Chong-Hendry form of forecast encompassing)

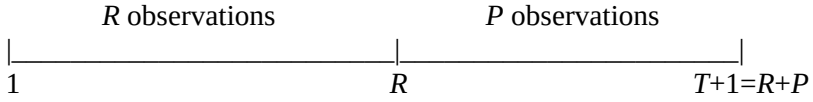
(3.6) $Ee_{1t}Z_t'=0$. (Chao, Corradi and Swanson test).

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

(3.7) $\gamma^* \neq 0$, $Ee_{1t}^2-Ee_{2t}^2>0$, $Ee_{1t}(X_{2t}'\beta_2^*)>0$, $-Ee_{1t}(X_{1t}'\beta_1^*- X_{2t}'\beta_2^*)>0$, $Ee_{1t}Z_t' \neq 0$.

That $Ee_{1t}^2-Ee_{2t}^2$ is positive, and that $Ee_{1t}Z_t'$ is nonzero, when $\gamma^* \neq 0$, is evident. That the correlation between e_{1t} and $X_{2t}'\beta_2^*$ is positive is most easily seen in the special case in which $EX_{1t}Z_t'=0$ (i.e., X_{1t} and Z_t are orthogonal). For in this case, $\beta_1^*=\delta^*$, and a little algebra establishes that $Ee_{1t}(X_{2t}'\beta_2^*) = E(Z_t'\gamma^*)^2$, which is positive. More generally, if we let $\tilde{Z}_t$ denote the residual of the projection of Z_t onto X_{1t}, $\tilde{Z}_t = Z_t-E(Z_t|X_t)$, then $Ee_{1t}(X_{2t}'\beta_2^*) = E(\tilde{Z}_t'\gamma^*)^2>0$. (To prevent confusion, we repeat that when $\gamma^* \neq 0$, then, in general, the model 2 coefficient vector on X_{1t} is not β_1^*.) Since $Ee_{1t}(X_{1t}'\beta_1^*)=0$ even under the alternative, it also follows that $-Ee_{1t}(X_{1t}'\beta_1^*- X_{2t}'\beta_2^*)>0$.

One uses out of sample prediction errors to form sample analogues of the moments in (3.4), (3.5) and (3.6). To state precisely how one might do so requires some extra notation.[3] 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 $R+P=T+1$. Schematically:

(3.8)


Let $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$ denote least squares estimates that rely on data from period t or earlier. We distinguish two schemes for using data to construct the regression estimates, because asymptotic and finite sample results differ for the two. 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 observation R+1 (recall that we are assuming one step ahead predictions, for simplicity); one then estimates β_1^* and β_2^* with data from 1 to R+1, with the new estimate used to predict observation R+2; ....; finally, one estimates β_1^* and β_2^* with data from 1 to T, with the final estimate used to predict observation T+1. 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, ..., the final with a sample running from T-R+1 to T. 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). The rolling scheme is relatively attractive when one wishes to guard against moment or parameter drift that is difficult to model explicitly.

It may help to illustrate with a simple example. Suppose model 1 is a univariate zero mean AR(1): $X_{1t}=y_{t-1}$, $y_t=\beta_1^* y_{t-1}+e_{1t}$. Then the sequence of P estimates of β_1^* are generated as follows for $t=R, \ldots, T$:

(3.9) recursive: $\hat{\beta}_{1t}=[\sum_{s=1}^{t}(y_{s-1}^2)]^{-1}[\sum_{s=1}^{t} y_{s-1}y_s]$;

rolling: $\hat{\beta}_{1t}=[\sum_{s=t-R+1}^{t}(y_{s-1}^2)]^{-1}[\sum_{s=t-R+1}^{t} y_{s-1}y_s]$.

In each case, the one step ahead prediction error is $\hat{e}_{1t+1} \equiv y_{t+1}-y_t\hat{\beta}_{1t} \equiv y_{t+1}-X_{1t+1}\hat{\beta}_{1t}$. (Note the dating: X_{1t+1}, not X_{1t}.)

More generally, write the predictions and prediction errors as

(3.10) $\hat{y}_{1t+1} \equiv X_{1t+1}'\hat{\beta}_{1t}$, $\hat{e}_{1t+1} \equiv y_{t+1}-\hat{y}_{1t+1}$, $\hat{y}_{2t+1} \equiv X_{2t+1}'\hat{\beta}_{2t}$, $\hat{e}_{2t+1} \equiv y_{t+1}-\hat{y}_{2t+1}$.

(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 sample analogues that may be used to test (3.4) to (3.6), together with the acronyms that are used to reference these in the table are:

(3.11) $P^{-1}\sum_{t=R}^{T}\hat{e}^2_{1t+1} - P^{-1}\sum_{t=R}^{T}\hat{e}^2_{2t+1} \equiv \hat{\sigma}^2_1-\hat{\sigma}^2_2$, (MSPE-normal)

(3.12) $P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(X_{2t+1}'\hat{\beta}_{2t})$ (CH)

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

CH may be clearer if we explicitly note that in the notation of (3.10), CH is $P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}\hat{y}_{2t+1}$.

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

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

Thus MSPE-normal may be written

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

Under the null, e_{1t} is uncorrelated with both X_{1t} and X_{2t}. It seems reasonable to expect, then, that $P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) \approx 0$ (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 $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2 <0$, *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 $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$. As in Clark and West (2005), we call this *MSPE-adjusted*:

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

We see from (3.14) that MSPE-adjusted is

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

Thus, under the alternative we expect MSPE-adjusted to be positive, since, as stated in (3.7), under the alternative $-Ee_{1t}(X_{1t}'\beta_1^* - X_{2t}'\beta_2^*)>0$. Hence we use one tailed tests in our simulations and empirical examples.

We shall compare, via simulations, the performance of MSPE-normal (3.11), Chong-Hendry (3.12), Chao et al. (3.13), and MSPE-adjusted (3.15). For each statistic, we rely on heteroskedasticity and autocorrelation consistent variance-covariance matrices. For CH and CCS, we adjust these variance covariance matrices for the reliance of predictions on estimated regression parameters as recommended by West and McCracken (1998) and Chao et al. (2001). We then compute t-statistics (MSPE-normal, CH, MSPE-adjusted) or chi-squared statistics (CCS) and use standard critical values. Some details are presented in a subsequent section. We wish to discuss here theoretical appropriateness of use of standard critical values. Unless otherwise stated, we maintain stationarity and moment conditions of the sort spelled out in detail in West (1996) or Giacomini and White (2004).

For all four statistics, standard critical values are appropriate when $P/R\rightarrow 0$ under an asymptotic approximation in which $R\rightarrow\infty$, $P\rightarrow\infty$ (West (1996), McCracken (2004), Clark and McCracken (2001, 2005a)). In many applications P is small relative to R but not so small as to make $P/R \rightarrow 0$ obviously attractive, an inference supported by simulation results reported below.

So we consider the complications that result if the $P/R\rightarrow 0$ condition seems unappealing. Let us take each of our four statistics in turn, with the discussion of MSPE-adjusted sufficiently involved that we

put it in a separate section. Throughout we rule out $P/R \to 0$, assume $P \to \infty$, and maintain stationarity and moment conditions of the sort spelled out in detail in West (1996) or Giacomini and White (2004). We do not attempt to maintain a uniform set of conditions for asymptotic analysis. In particular, as will be clear, we will freely move between approximations that result when $R \to \infty$ as $P \to \infty$ and those that result when R is held fixed as $P \to \infty$, opportunistically relying on whichever seems to give better guidance to our finite sample results. Approximations for $R \to \infty$ are available for both rolling and recursive schemes, while R fixed requires the rolling scheme.

•MSPE-normal: To our knowledge there is no appealing set of conditions under which the t-statistics computed using MSPE-normal are asymptotically normal. The presence of the negative term $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$ 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. McCracken (2004) derives a non-standard distribution that results under an $R \to \infty$ approximation. We shall find it useful to interpret certain simulation results for MSPE with an R fixed approximation.

•CH: Conditions that establish asymptotic normality, once one adjusts for sampling error in estimation of regression parameters used to make predictions, may be found in West (1996), West and McCracken (1998) and West (2005). These conditions include $R \to \infty$. For the rolling scheme, one must divide the usual t-statistic by a certain function of R and P to produce an asymptotically normal test statistic. This function is spelled out in a subsequent section. For the recursive scheme, CH requires $\beta_1^* \neq 0$ for asymptotic normality.

•CCS: See Chao et al. (2001) for conditions that establish asymptotic normality. These conditions include $R \to \infty$. One has to adjust for sampling error in estimation of regression parameters used to make predictions, as described in Chao et al. (2001).

4. INFERENCE ON MSPE-ADJUSTED

With a little algebra, it can be established that

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

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.[4]

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. (To illustrate these terms, consider the univariate example of forecasting $y_{t+\tau}$ using y_t. The "direct" method estimates $y_{t+\tau} = y_t\gamma + u_{t+\tau}$ by least squares, uses $y_t\hat{\gamma}_t$ to forecast, and computes a sample average of $(y_{t+\tau}-y_t\hat{\gamma}_t)^2$. The "iterated" method estimates $y_{t+1} = y_t\beta + e_{t+1}$, uses $y_t(\hat{\beta}_t)^\tau$ to forecast, and computes a sample average of $[y_{t+\tau}-y_t(\hat{\beta}_t)^\tau]^2$.[5])

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 functionals of Brownian motion that do not depend on the specifics of the DGP. The functionals do 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≤dimension of

$Z_t \leq 20$ and for 20 different limiting values of P/R (specifically, P/R = 0.1, 0.2, 0.4, 0.6, 0.8., 1.0, 1.2, 1.4, 1.6, 1.8., 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 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 recursive schemes,

(4.2) .90 quantile $\leq$ 1.282 $\leq$ .95 quantile

Recall that for one-tailed tests using standard normal critical values, the .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 .10 tests that have actual size somewhere between .05 and .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 .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 .90 and .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 .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 the 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 .05 and .10.

As well, it is possible that the critical values for values of P/R not tabulated strongly violate the inequalities. For example, consider the .90 and .95 quantiles for P/R=1.0 and P/R=1.2, dimension of Z_t =4 (as in one of our DGPs used in the simulation and in our empirical example), rolling scheme. These are:

	.90 quantile	.95 quantile
P/R=1.0	1.10	1.46
P/R=1.2	1.10	1.47

It is possible that for some value of P/R lying between 1.0 and 1.2, the .90 quantile shoots well above 1.282, or the 0.95 value drops well below. 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 .05 and.10, when the dimension of $Z_t \leq 20$ and for $P/R \leq 20.0$. It might be of interest to note when the .90 quantile is closer rather than farther from 1.282, to guide when inference is likely to be relatively accurate. As a rule, the .90 quantile is relatively near to 1.282, and thus tests using 1.282 as the critical value are likely to have size relatively near .10, under one or more of the following

circumstances:

•P/R near 0 (recall that as $P/R \to 0$, MSPE-adjusted becomes asymptotically, normal, so the .90 quantile approaches 1.282 as $P/R \to 0$);

•larger dimensions of Z_t;

•rolling rather than recursive;

•for rolling, for $P/R>1$. (The quantiles for rolling are broadly U-shaped in P/R, initially falling as P/R increases from 0.1 but then rising as P/R approaches 20.0. The terminal (P/R=20.0) values generally are below but near the initial (P/R=0.1) values. For the recursive scheme, the quantiles broadly fall as P/R increases, with terminal values below those for the rolling scheme.)

Recall that the .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

(4.3) .95 quantile.$\leq$ 1.645 $\leq$ .99 quantile.

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 .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 .01 and .05 (approximately), for P and R sufficiently large. We focus on the 1.282 critical value in part because our sense is that practical implications of .10 and .05 rejections are more similar than those of .05 and .01 rejections, but mostly because it seems that tests of size .05 to .10 are of more interest to applied researchers than tests of size .01 to .05. Of course, others might have a different view, in which case use of a 1.645 critical value will be of interest.

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, we tentatively offer an interpretation of use of standard critical values. This argument follows Clark and West (2005), who in turn follow Giacomini and White (2004). We do not present a formal argument for using standard critical values with the recursive scheme outside the environment of the previous paragraph. For what it is worth, the simulations below suggest the normal approximation can work okay for the recursive scheme even outside this environment, though we have no evidence or argument that these simulations are representative.

For the rolling scheme, consider an asymptotic approximation in which R is held fixed, and $P\rightarrow\infty$. Giacomini and White (2004) show that under suitable conditions, $\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$ is a well behaved random variable. These "suitable conditions" relax some of the Clark and McCracken (2001, 2005a) restrictions: general nonlinear parametric models and estimators are allowed,, and multistep forecasts may be made with the iterated as well as the direct method. The result is that $\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$ obeys the usual law of large numbers and central limit theorem as $P\rightarrow\infty$:

$$(4.4)\quad -2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) \rightarrow_p -2E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}),$$

$$-2P^{-1/2}[\sum_{t=R}^{T}\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) - E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})] \sim_A N(0,V),$$

$$V = 4 \times \text{long run variance of } \hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}).$$

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

Consider first the Clark and West (2005) environment in which $\beta_1^*=0$ (i.e., the null model is that y_{t+1} is a martingale difference and so one always predicts that y_{t+1} will be zero), and the rolling scheme is used for prediction. Then $\hat{e}_{1t+1}=e_{1t+1}$, $\hat{y}_{1t+1}=0$ and MSPE adjusted = $-2P^{-1}\sum_{t=R}^{T}e_{1t+1}\hat{y}_{2t+1}$. Since we take R as fixed, $e_{1t+1}\hat{y}_{2t+1}$ is a stationary random variable. In this special case, the expectation of MSPE-adjusted is zero. Standard normal inference will yield accurately sized tests for P sufficiently large.

With R fixed, this result clearly may not hold when the null model relies on estimated parameters. We have $\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) = e_{t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1}) + X_{1t+1}'(\beta_1^*-\hat{\beta}_{1t})(\hat{y}_{1t+1}-\hat{y}_{2t+1})$. The first term has expectation zero, but the second term, in general, does not. Thus in (4.4), the value of $E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$ in general is non zero. So under the null given in (3.3), as well as under the alternative given in (3.7), MSPE-adjusted will converge in probability to a nonzero value as $P\rightarrow\infty$ 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) .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.

While this is the only formally established interpretation we can offer to application of our tests when the dimension of Z_t is more than 1, and there are conditionally heteroskedastic and/or multistep forecasts, we leave the door open to interpreting those simulations as hypothesis tests. We find that the simulated critical values are not too far from standard normal critical values (though we recognize the possibility that there may be DGPs for which the critical values are quite distant from normal ones). As well, observe in (4.4) that the value of $E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$ depends not only on the DGP but also on the fixed value of R. Larger values of R imply smaller absolute values of $E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$; for R arbitrarily large, $E\hat{e}_{1t+1}(\hat{y}_{1t+1}-\hat{y}_{2t+1})$ will be arbitrarily near zero. So for R reasonably large, we expect standard normal critical values to be reasonably accurate, for any DGP.

Earlier, we observed that an approximation in which $P/R \rightarrow 0$ is not obviously appealing. The approximation that we have just discussed, which holds R fixed as $P\rightarrow\infty$, thereby implying $R/P \rightarrow 0$, also may not be obviously appealing. Nonetheless, our simulation evidence finds that the R fixed approximation works better than the $P/R \rightarrow 0$ 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, $P/R \geq 2$); the $P/R \rightarrow 0$ approximation

rationalizes the behavior of MSPE-normal (theoretically approximately normal) only for small and empirically uncommon values of P/R (say, $P/R \leq .10$).

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 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.*"[8]

5. TEST STATISTICS

Our simulations allow for possibly multiperiod predictions. These are constructed from regressions using overlapping data, as described in the next section. With a prediction horizon of $\tau \geq 1$, let $\hat{e}_{1t+\tau}$ and $\hat{e}_{2t+\tau}$ be the multistep forecast errors. For MSPE-normal, CH, CCS, or MSPE-adjusted, let $\hat{f}_{t+\tau}$ be an observation on the object of interest, with sample average $\bar{f}$. For example, for MSPE-normal,

$$\hat{f}_{t+\tau} = \hat{e}^2_{1t+\tau} - \hat{e}^2_{2t+\tau}, \quad \bar{f} = (P-\tau+1)^{-1}\sum_{t=R+\tau-1}^{T}\hat{e}^2_{1t+\tau} - (P-\tau+1)^{-1}\sum_{t=R+\tau-1}^{T}\hat{e}^2_{2t+\tau} \equiv (P-\tau+1)^{-1}\sum_{t=R+\tau-1}^{T}\hat{f}_{t+\tau}.$$

For CH, recursive scheme, and MSPE-normal and MSPE-adjusted rolling and recursive schemes, our test statistic is

(5.1) $\sqrt{P}\bar{f}$ / [estimator of long run variance of $\hat{f}_{t+\tau}-\bar{f}$]$^{1/2}$.

For rolling regressions, CH was adjusted for uncertainty due to estimation of regression parameters as suggested in West and McCracken (1998). Let $\lambda=1-(P^2/3R^2)$ if $P\leq R$, $\lambda=(2R/3P)$ if $P>R$. Let $\bar{f}$ be the sample value of CH, $\bar{f} = P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+\tau}(X_{2t+1}'\hat{\beta}_{2t})$. Then for such $\bar{f}$, scaling the denominator of the usual t-statistic (5.1) by λ produces an asymptotically normal test statistic:

(5.2) $\sqrt{P}\bar{f} / [\lambda \times$ estimator of long run variance of $\hat{f}_{t+\tau}-\bar{f}]^{1/2}$.

CCS was adjusted for uncertainty due to estimation of regression parameters as described in Chao et al. (2001).

In estimation of the long run variance, for one step ahead predictions we used the sample variance (MSPE-normal, MSPE-adjusted and CH) or the usual heteroskedasticity consistent estimator (CCS). For multistep predictions of horizon τ, we used Newey and West (1987) with a bandwidth of 1.5τ.

For MSPE-adjusted, with simulation critical values, we followed the procedure described in Clark and McCracken (2005a).

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 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) and Tauchen (2001)):

(6.1) $y_t = 0.5 + \gamma^* z_{t-1} + e_{1t}$, $X_{1t}=1$, $X_{2t}=(1, z_{t-1})'$, $z_t = 0.15 + 0.95 z_{t-1} + v_t$,

$E_{t-1} e_{1t}=0$, $E_{t-1} v_t=0$, $var(e_{1t})=18.0$, $\text{var}(v_t)=0.025$, $corr(e_{1t},v_t)=-0.75$;

$\gamma^*=0$ in experiments evaluating size, $\gamma^*=0.35$ in experiments evaluating power.

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:

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

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

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

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.

We consider forecasts for various horizons, following the common approach of using overlapping data to make a τ-step ahead forecast of $y_{t+\tau,t} = y_{t+\tau} + y_{t+\tau-1} + \ldots + y_{t+1}$. (In this notation, $y_{t+1,1} = y_{t+1}$). The forecasts are constructed from least squares regressions of the following forms:

(6.4) $y_{t+\tau,t} = \beta_1^* + e_{1t+\tau} \equiv X_{1t}\beta_1^* + e_{1t+\tau}$, (null model)

$y_{t+\tau,t} = \delta^* + \gamma^* z_t + e_{2t+\tau} \equiv X_{2t}'\beta_2^* + e_{2t+\tau}$. (alternative model)

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) and 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:

$$(6.5) \quad y_t = 2.237 + .261y_{t-1} + \gamma_1^* z_{t-1} + \gamma_2^* z_{t-2} + \gamma_3^* z_{t-3} + \gamma_4^* z_{t-4} + e_{1t},$$

$$z_t = .804z_{t-1} - .221z_{t-2} + .226z_{t-3} - .205z_{t-4} + v_t,$$

$$\text{var}(e_{1t}) = 10.505, \text{var}(v_t) = .366, \text{cov}(e_{1t}, v_t) = 1.036,$$

$$\gamma_i^* = 0, \; i = 1,..., 4, \text{ in size experiments;}$$

$$\gamma_1^* = 3.363, \; \gamma_2^* = -.633, \; \gamma_3^* = -.377, \; \gamma_4^* = -.529 \text{ in power experiments.}$$

The forecasting models for the τ-step ahead forecast of $y_{t+\tau,t} = y_{t+\tau} + y_{t+\tau-1} + ... + y_{t+1}$ are

$$(6.6) \quad y_{t+\tau,t} = \beta_{11}^* + \beta_{12}^* y_t + e_{1t+\tau} \equiv X_{1t}\beta_1^* + e_{1t+\tau}, \qquad \text{(null model)}$$

$$y_{t+\tau,t} = \delta_1^* + \delta_2^* y_t + \gamma_1^* z_t + \gamma_2^* z_{t-1} + \gamma_3^* z_{t-2} + \gamma_4^* z_{t-3} + e_{2t+\tau}$$

$$X_{1t}'\delta^* + Z_t'\gamma^* + e_{2t+\tau}. \equiv X_{2t}'\beta_2^* + e_{2t+\tau}. \qquad \text{(alternative model)}$$

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+ τ through R+P, with the total number of predictions being P- τ +1. The number of simulations is 10,000.

Throughout, we present results for CH, MSPE-normal and MSPE-adjusted where "rejection" is defined as: the t-statistic is greater than +1.282. For CCS, we refer to the .90 quantiles of a $\chi^2(1)$ (DGP 1) or $\chi^2(4)$ (DGP 2) distribution. For CH and CCS, this defines a test of nominal size .10. The critical values in McCracken (2004) indicate that for MSPE-normal and conditionally homoskedastic disturbances, this defines a test of nominal size below 0.10, typically well below 0.10. Interpretation for MSPE-adjusted was presented above and will be reviewed below. For MSPE-adj. simul. cvs., we define rejection as: the t-statistic is above the .90 quantile in the simulated distribution. This defines a test of nominal size .10. An Appendix available on request from the authors contains results when we use a standard .05 cutoff (e.g., t-statistic cutoff of +1.645). We summarize below some results from that Appendix.

6.2 Simulation Results: One Step Ahead Forecasts

In this section, we consider one step ahead forecasts. As discussed above, for MSPE-adjusted, our rejection rule defines a test of size between .05 to .10, where the size depends on the sampling scheme, dimension of Z_t and P/R.

Tables 1 and 2 present results for MSPE-adjusted and MSPE-normal. Table 1 considers conditionally homoskedastic disturbances, while Table 2 allows conditional heteroskedasticity for DGP 1. Table 1 contains results for DGP 1, R=120 and R=240 and for DGP 2, R=80 and R=120. In Table 2, results for DGP 1, R=120 are presented.

In both tables, and for both DGPs in Table 1, the results for MSPE-adjusted are in good conformity with the asymptotic analysis presented above. Most notably, actual sizes generally fall between .05 and .10. The only exceptions are in Table 2, panels A2 and B2 for smaller sample sizes of

P=120 and P=240. As well sizes tend to be relatively close to .10 in ways that are consistent with that analysis. In Table 1, sizes are closer to .10 than to .05 for rolling rather than recursive and for larger rather than smaller dimension of Z_t (DGP 2 rather than DGP 1). A tendency for size to first fall and then rise with P/R is seen in about half the entries (panels A2, A3 and B3 in Table 1, panel A1 in Table 2). Our findings are consistent with Clark and McCracken's (2001, 2005a) results for their Enc-t statistic.[9]

As in Clark and McCracken (2001, 2005a), Clark and West (2005) and Corradi and Swanson (2005), MSPE-normal is seriously undersized. In Table 1, the median size is .006 in DGP 1 and .003 in DGP 2. 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 $\hat{\sigma}_1^2-(\hat{\sigma}_2^2\text{-adj.}) \equiv \hat{\sigma}_1^2-(\hat{\sigma}_2^2 - P^{-1}\sum_{t=R}^{T}[\hat{y}_{1t+1}-\hat{y}_{2t+1}]^2)$. (See (3.15).) To illustrate the mean and median bias in MSPE normal, consider DGP 1, with conditionally homoskedastic disturbances, R=120 and P=720 (Table 1, panel 1A). Across 10,000 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 $P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$ 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 $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$ on the r.h.s. of (3.14)). 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 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 $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$ 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 $R\rightarrow\infty$, $\hat{y}_{1t+1}-\hat{y}_{2t+1} \rightarrow_p 0$), thus explaining the improved size with bigger R.

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 .10, does credibly against other competitors. Table 3 presents results for CH, CCS and MSPE-adj. simul. cvs for the recursive scheme. Results for MSPE-adjusted are repeated from Tables 1 and 2, to facilitate comparison. We report the recursive scheme in detail to be conservative; results for the rolling scheme, which are reported in the Appendix, are slightly more supportive for MSPE-adjusted, as one might guess by comparing panels A and B in Table 1.

We see in Table 3, panels A1, A2, B1 and B2, that in DGP 1 the Chong-Hendry test statistic is also undersized, though not as seriously as is MSPE-normal. CH is better sized in DGP 2 (panels A3 and A4). The CCS statistic is a bit oversized in DGP B (panels A3 and A4), but is very nicely sized in DGP A, even in conditionally heteroskedastic DGPs (panels 1 and B2) . MSPE with simulation-based critical values is slightly oversized in all DGPs, especially in the presence of multiplicative conditional heteroskedasticity (panels B1 and B2).

The most glaring discrepancy between our asymptotics and finite sample performance is for CH. We therefore experimented with some larger sample sizes to see what sizes were required to have the asymptotic approximation for CH work tolerably well. We set $P/R = 1$ and experimented with increasingly larger values of P, using DGP 1. Results, with value of P (=value of R) in parentheses, and with the results for P=120 and P=240 repeated from Table 3: .040 (120), .050 (240), .078 (1000), .092 (3000). At this point we stopped. It is clear that very large sample sizes are required for the asymptotic approximation to work reasonably well. We do not know why CH requires unusually large samples.

For the rolling scheme, performance for CCS and MSPE-adj. simul. cvs was qualitatively similar; performance for CH degraded substantially. (Details are in the appendix.) Perhaps a good summary statistic to compare the five test statistics (the four in Table 3, and MSPE-normal) is the median empirical size. Across all 48 DGPs–the 24 given in Table 3, and the 24 additional ones for the rolling scheme–median empirical sizes were: MPSE-adjusted: .08; MSPE-normal: .01; CH: .05; CCS: .11;

MSPE-adj. simul. cvs: .11.

Figures 1 and 2 present smoothed density estimates of the test statistics. In Figure 1, results correspond to values for DGP 1 in Table 1, panel B2, and Table 3, panel A2, for P=120, 240 and 720. (P=360 was omitted from the figure for legibility.) Figure 2 presents comparable figures for the rolling scheme. In Figure 2, results for MSPE-adjusted and MSPE normal correspond to values reported in Table 1, panel A2; results for CH and CCS are reported in the Appendix. While results in Table 3 for CCS rely on chi-squared statistics, we plot the square root of that statistic in the Figures.

That MSPE-adjusted, MSPE and CH are undersized in both Figures is clear: 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.[10] In either figure, a comparison of panels C and D reveals clearly that MSPE-adjusted will be better sized than MSPE-normal, because of the sharper leftward shift in MSPE-normal. The distribution of MSPE-normal is piling up on what we called $\hat{y}(R)$ as P increases. For MSPE-adjusted, Figure 2D shows that for the rolling scheme, undersizing diminishes as P increases, consistent with the quantiles in Clark and McCracken (2001). The poor performance of *CH* under the rolling scheme is clearly reflected in the densities in Figure 2A. Finally, panel B in both figures illustrates the good performance of CCS.

Results for tests using a critical value of +1.645 are presented in the not for publication Appendix. They tell the same story. For MSPE-adjusted, of 48 sets of simulations, 44 had size between 0.01 and 0.05, with 4 (all involving multiplicative conditional heteroskedasticity) slightly greater than 0.05. The median size was 0.04. Median sizes for other test statistics were: MSPE-normal: 0.00; CH: 0.03; CCS: 0.05; MSPE-adj. simul. cvs..: 0.06.

Table 4 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 both the “MSPE-adjusted” and “MSPE-adj. simul. cvs” entries in Table 1 because size adjusted power is identical for the two.

In DGP 1, size adjusted power is best for MSPE-adjusted, worst for CCS, with MSPE-normal and CH in the middle. In DGP 2, power is best for MSPE-adjusted, worst for CH, with MSPE-normal and CCS 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. Table 5 presents unadjusted power - that is, power that results if one uses the asymptotic normal critical value of 1.282 (or, for CCS, 2.71 for DGP 1 or 7.78 for DGP 2). MSPE-adjusted, with simulation critical values, has modestly better power than MSPE adjusted. The other three tests have distinctly poorer power.

Inspection of the numbers in Panels A and B of Table 5 indicates that for DGP 1, even the best unadjusted power may not be very good. The very largest figure in the table is 0.394 (panel B, P=720, MSPE-adj. simul. cvs). 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% lower than that of the null model (i.e., the R^2 in the alternative model is about .05). With such a small amount of predictability, it will take many, many observations to have high probability of rejecting the null.

In any event, we conclude that of the four statistics that do not require simulations to compute critical values, MSPE-adjusted has the best power.

6.3 Simulation Results: Multistep Ahead Forecasts

Table 6 presents results for multistep horizons. The DGPs are as in Table 1 (conditionally homoskedastic disturbances). The forecast horizon τ is set to 12 in our pseudo-asset pricing DGP 1, consistent with a one year horizon for monthly data; the horizon τ is set to 4 in our pseudo-macro DGP 2 to match a one year horizon for quarterly data. Since the dimension of Z_t is 1 in DGP 1, for large enough P and R, the size of tests of MSPE-adjusted will fall between .05 and .10. Since the dimension of Z_t is 4 in DGP 2, we cannot make such a statement.

But in practice, there are no qualitative differences between the simulation results for the two

DGPs. All the statistics have difficulty when *P* is relatively small. This likely reflects the fact that our smallest values of *P* include fewer than 10 nonoverlapping sets of forecasts (though, as explained above, we do use all 109 (=120-12+1, DGP 1) or 37 (=40-4+1, DGP 2) forecasts in computing the statistics). With so few observations, estimation of the variance-covariance matrix is difficult. It seems that these difficulties inflate the test statistics. For MSPE-normal, and for small *P*, this results in modest rather than serious undersizing (e.g., size of .061 in panel A, versus the.012 figure in panel A of Table 1): the inflation from mis-estimation of the covariance matrix partially offsets the miscentering that results from failure to account for noise in the alternative model's forecasts. A similar phenomenon explains the relatively good size of CH for small *P*. As *P* increases, size improves. It seems that MSPE-adjusted is the most reliable statistic.

That MSPE-adjusted performs better than MSPE-normal once *P* is large enough to permit reasonably accurate estimation of the relevant standard error again reflects better centering of MSPE-adjusted. For DGP 1, *R*=120, *P*=720, for example, across the 10,000 simulations we have: mean of numerator of t-statistic for MSPE-normal = mean of $\hat{\sigma}_1^2$-$\hat{\sigma}_2^2$ = -35.8; mean of numerator of t-statistic for MSPE-adjusted = mean of $\hat{\sigma}_1^2$-($\hat{\sigma}_2^2$-adj.) = -1.8. (To scale these figures, note that the variance of the MA(11) forecast error is 11×18=196.)

7. EMPIRICAL EXAMPLE

To illustrate our approach, we apply the MSPE-adjusted, MSPE-normal, CH, 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 (references in addition to those already cited include Fama and French (1988) and Pesaran and Timmermann (1995)), 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-squared.

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 U.S. 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 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 7 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 .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 2.07 for both rolling and recursive forecasts, allowing rejection at a significance level between 0.01 and 0.05 (see equation (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 $P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$. The adjustment is .67 or .68 for stock return forecasts (corresponding to about 3 to 4 percent of the alternative model's MSPE) and 1.01 to 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 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 adust 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 a modestly undersized tests: one-sided tests using 1.282 as the critical value will, in large samples, have actual size somewhere between .05 and .10; one sided tests using 1.645 will have size between .01 and .05. Simulations support our recommended procedure.

FOOTNOTES

1. Recent examples include Lettau and Ludvigson (2001), Guo (2002), Goyal and Welch (2003), Ang et al. (2004), and Campbell and Thompson (2005).

2. References in addition to those already given include Lettau and Ludvigson (2001), Stock and Watson (2002, 2003, 2004), Goyal and Welch (2003), Marcellino et al. (2003), Diebold and Li (2004), Orphanides and van Norden (2005), Rapach and Weber (2004), Clark and McCracken (2005b) and Shintani (2005).

3. The prose in the following two paragraphs is a lightly edited version of a passage in West (2005).

4. 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.

5. Of course, if the AR(1) model for y_t is correct, then $\gamma=\beta^{\tau}$ and $u_{t+\tau}=e_{t+\tau}+\beta\, e_{t+\tau-1}+...+\beta^{\tau-1}e_{t+1}$; the two forecasts may differ, even in a large sample, if the AR(1) model is incorrect. See Ing (2003) and Marcellino et al. (2004) for theoretical and empirical comparison of direct and iterated methods.

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.

7. Giacomini and White (2004) propose what they call an unconditional test of the equality of the raw MSPE difference. They similarly state that the long run variance must be computed even for one step ahead forecasts. Their analysis of the raw MSPE difference departs from ours in that they seem to maintain the assumption that the raw MSPE difference is centered at zero, while we conclude that the difference is shifted downwards, see the discussion below (3.14).

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

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

10. West and McCracken (1998) also found poor finite sample performance for CH, with larger distortions occurring for larger P. But in West and McCracken (1998), CH was oversized. The figure helps explain why we instead find CH undersized. West and McCracken (1998) used two tailed tests, we use one tailed tests. It is clear from the figure that with two tailed tests, CH is increasingly oversized as P increases.

REFERENCES

Ang, Andrew, Monika Piazzesi, and Min Wei, 2004, "What does the Yield Curve Tell us about GDP Growth?", forthcoming, *Journal of Econometrics*.

Campbell, John Y., 2001, "Why Long Horizons? A Study of Power Against Persistent Alternatives," *Journal of Empirical Finance* 8, 459-91.

Campbell, John Y., and Samuel B. Thompson, 2005, "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?", manuscript, Harvard University.

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

Chong, Y.Y. and David F. Hendry, 1986, "Econometric evaluation of linear macro-economic models", *Review of Economic Studies*, 53, 671-690.

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

Clark, Todd E. and Michael W. McCracken, 2005a, "Evaluating Direct Multistep Forecasts," manuscript, Federal Reserve Bank of Kansas City.

Clark, Todd E., and Michael W. McCracken, 2005b, "The Predictive Content of the Output Gap for Inflation: Resolving In-Sample and Out-of-Sample Evidence," forthcoming, *Journal of Money, Credit, and Banking*.

Clark, Todd E. and Kenneth D. West, 2005, "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis,"forthcoming, *Journal of Econometrics*.

Cooper, Michael, Roberto C. Gutierrez Jr., and William Marcum, 2005, "On the Predictability of Stock Returns in Real Time," *Journal of Business*, 469-500.

Corradi, Valentini and Norman R. Swanson, 2005, "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," manuscript, Rutgers University.

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

Diebold, Francis X. and Robert S. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Fama, Eugene F., and Kenneth R. French, 1988, "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics* 22, 3-25.

Faust, Jon, John H. Rogers and Jonathan H. Wright, 2005, "News and Noise in G-7 GDP Announcements," *Journal of Money, Credit and Banking* 37, 403-420.

Giacomini, Rafaella and Halbert White, 2004, "Tests of Conditional Predictive Ability," manuscript, University of California at San Diego.

Goyal, Amit and Ivo Welch, 2003, "Predicting the Equity Premium With Dividend Ratios," *Management Science* 49, 639-654.

Guo, Hui, 2002, "On the Out-of-Sample Predictability of Stock Market Returns," forthcoming, *Journal of Business*.

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

Ing, Ching-Kang, 2003, "Multistep Prediction in Autoregressive Processes," *Econometric Theory* 19, 254-279.

Inoue, Atsushi, and Lutz Kilian, 2004, "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," forthcoming, *Econometric Reviews*.

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

Lettau, Martin, and Sydney Ludvigson, 2001, "Consumption, Wealth, and Expected Stock Returns," *Journal of Finance* 56, 815-849.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson, 2003, "Macroeconomic Forecasting in the Euro Area: Country-Specific Versus Area-Wide Information," *European Economic Review* 47, 1-18.

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

Mark, Nelson, 1995, "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review* 85, 201-218.

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

Nelson, Charles R., and Myung J. Kim, 1993, "Predictable Stock Returns: The Role of Small Sample Bias," *Journal of Finance* 48, 641-61.

Newey, Whitney K. and Kenneth D. West, 1987, "A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703-708.

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

Pesaran, M. Hashem, and Allan Timmermann, 1995, "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance* 50, 1201-1228.

Rapach, David E., and Christian E. Weber, 2004, "Financial Variables and the Simulated Out-of-Sample Forecastability of U.S. Output Growth Since 1985:An Encompassing Approach," *Economic Inquiry* 42, 717-38.

Shintani, Mototsugu, 2005, "Nonlinear Forecasting Analysis Using Diffusion Indexes: An Application to

Japan," *Journal of Money, Credit, and Banking* 37, 517-538.

Stambaugh, Robert F., 1999, "Predictive Regressions," *Journal of Financial Economics* 54, 375-421.

Stock, James H., and Mark W. Watson, 2002, "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics* 20, 147-162.

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

Stock, James H., and Mark W. Watson, 2004, "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting* 23, 405-430.

Tauchen, George, 2001, "The Bias of Tests for a Risk Premium in Forward Exchange Rates," *Journal of Empirical Finance* 8, 695-704.

West, Kenneth D., 1996, "Asymptotic Inference About Predictive Ability," *Econometrica* 64 , 1067-1084.

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

West, Kenneth D., 2005, "Forecast Evaluation," manuscript, University of Wisconsin.

West, Kenneth D. and Michael W. McCracken, "Regression Based Tests of Predictive Ability," *International Economic Review* 39, 817-840.

Table 1
Empirical Size: 1-Step Ahead Forecasts

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	0.072	0.073	0.074	0.091	0.073	0.069	0.066	0.074
MSPE-normal	0.012	0.003	0.001	0.000	0.031	0.013	0.006	0.002

	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	0.086	0.079	0.083	0.091	0.082	0.078	0.076
MSPE-normal	0.015	0.003	0.001	0.000	0.026	0.008	0.003	0.001

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	0.070	0.067	0.059	0.054	0.075	0.066	0.062	0.058
MSPE-normal	0.024	0.015	0.008	0.003	0.034	0.021	0.015	0.008

	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	0.090	0.081	0.076	0.079	0.093	0.082	0.078	0.073
MSPE-normal	0.019	0.008	0.004	0.002	0.030	0.012	0.008	0.006

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(6.5); the alternative model includes lags of an AR(4) variable z_t along with the lag of y_t, again with parameters given in (6.5). 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 (3.11); *MSPE–adjusted* adjusts the difference in mean squared prediction errors to account for the additional predictors in the alternative models, see (3.15). *t*-statistics are computed by dividing the point estimate by its standard deviation.

4. The number of simulations is 10,000. The table reports the fraction of simulations in which each test statistics 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, 717 test statistics were greater than 1.282. After rounding, this led to the figure of .072 given in the table.

5. For large *P* and *R*, MSPE-adjusted has size between .05 and .10, while MSPE-normal has size below .10.

Table 2
Empirical Size: DGP 1 with Heteroskedasticity
1-Step Ahead Forecasts, R = 120

A. Rolling Regressions

	1. *GARCH*				2. *Multiplicative*			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.080	0.076	0.082	0.083	0.119	0.103	0.091	0.081
MSPE-normal	0.017	0.003	0.001	0.000	0.019	0.004	0.002	0.000

B. Recursive Regressions

	1. *GARCH*				2. *Multiplicative*			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.082	0.072	0.065	0.058	0.114	0.095	0.081	0.069
MSPE-normal	0.029	0.013	0.008	0.004	0.033	0.015	0.012	0.004

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 equation (6.2). In panel B, the predictand y_{t+1} has conditional heteroskedasticity of the form given in equation (6.3), in which the conditional variance at t is a function of z^2_{t-1}.

Table 3
Empirical Size: Other Test Statistics, Recursive Regressions, 1 Step Ahead Forecasts

A.Conditionally homoskedastic disturbances

	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	0.070	0.067	0.059	0.054	0.075	0.066	0.062	0.058
CH	0.040	0.039	0.032	0.028	0.058	0.050	0.041	0.038
CCS	0.097	0.101	0.100	0.097	0.106	0.102	0.097	0.095
MSPE-adj.:simul. cvs	0.125	0.114	0.111	0.105	0.117	0.114	0.106	0.105

	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	0.090	0.081	0.076	0.079	0.093	0.082	0.078	0.073
CH	0.085	0.085	0.078	0.078	0.093	0.089	0.085	0.083
CCS	0.147	0.120	0.112	0.105	0.144	0.114	0.107	0.102
MSPE-adj.:simul. cvs	0.109	0.107	0.105	0.107	0.111	0.105	0.101	0.101

B. Conditionally heteroskedastic disturbances, DGP 1, R=120

	1. *GARCH*				**2. *Multiplicative***			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.082	0.072	0.065	0.058	0.114	0.095	0.081	0.069
CH	0.050	0.038	0.034	0.031	0.078	0.063	0.053	0.045
CCS	0.111	0.103	0.104	0.101	0.096	0.092	0.092	0.090
MSPE-adj.:simul. cvs	0.129	0.120	0.112	0.107	0.172	0.153	0.135	0.123

Notes:

1. See notes to Table 1. The values for MSPE-adjusted are repeated from Tables 1 and 2.

2. *CH* is the Chong-Hendry (1986) forecast encompassing statistic, see (3.12); *CCS* is the Chao et al. (2001) statistic testing whether model 1 forecasts are uncorrelated with the additional predictors in model 2, see (3.13); *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. For large *P* and *R*, all three statistics have nominal size .10.

Table 4
Size-Adjusted Power: 1-Step Ahead Recursive Forecasts
Size = 10%

	A. *DGP 1, R=120*				**B. *DGP 1, R=240***			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.181	0.226	0.265	0.355	0.197	0.239	0.284	0.382
MSPE-normal	0.172	0.203	0.232	0.315	0.183	0.220	0.253	0.328
CH	0.167	0.197	0.210	0.236	0.132	0.152	0.161	0.174
CCS	0.058	0.057	0.055	0.057	0.058	0.052	0.053	0.059
	C. *DGP 2, R=80*				**D. *DGP 2, R=120***			
	P=40	*P=80*	*P=120*	*P=160*	*P=40*	*P=80*	*P=120*	*P=160*
MSPE-adjusted	0.974	0.999	1.000	1.000	0.975	1.000	1.000	1.000
MSPE-normal	0.850	0.981	0.998	1.000	0.834	0.976	0.997	1.000
CH	0.497	0.714	0.852	0.915	0.485	0.725	0.855	0.925
CCS	0.591	0.925	0.992	0.999	0.582	0.924	0.992	0.999

Notes:

1. In panels A and B, the DGP is defined in equation 6.1, with the nonzero value of γ^* given in that equation. In panels C and D, the DGP is defined in (6.5), with nonzero values of γ_i^* given in (6.5). In each simulation, one step ahead forecasts of y_{t+1} are formed from rolling estimates of a regression of y_t on X_{1t} and on X_{2t}, for X_{1t} and X_{2t} defined in (6.4) and (6.6).

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" and "MSPE-adj. simul. cvs" use the same test statistic, size adjusted power is identical for the two.

Table 5
Unadjusted Power: 1-Step Ahead Recursive Forecasts

	A. *DGP 1, R=120*				**B. *DGP 1, R=240***			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.140	0.162	0.189	0.260	0.154	0.178	0.202	0.284
MSPE-normal	0.050	0.039	0.030	0.030	0.073	0.062	0.058	0.057
CH	0.068	0.073	0.075	0.076	0.076	0.071	0.069	0.080
CCS	0.056	0.058	0.055	0.054	0.061	0.054	0.051	0.056
MSPE-adj.:simul. cvs	0.215	0.252	0.282	0.366	0.220	0.267	0.297	0.394
	C. *DGP 2, R=80*				**D. *DGP 2, R=120***			
	P=40	*P=80*	*P=120*	*P=160*	*P=40*	*P=80*	*P=120*	*P=160*
MSPE-adjusted	0.969	0.999	1.000	1.000	0.972	1.000	1.000	1.000
MSPE-normal	0.575	0.823	0.935	0.980	0.617	0.848	0.944	0.983
CH	0.451	0.676	0.813	0.894	0.466	0.699	0.832	0.907
CCS	0.677	0.940	0.993	0.999	0.672	0.935	0.993	0.999
MSPE-adj.:simul. cvs	0.976	1.000	1.000	1.000	0.979	1.000	1.000	1.000

Notes:

1. This table differs from Table 4 in that power is computed using critical values also used in Tables 1 to 3.

Table 6
Empirical Size: Year-Ahead Forecasts

A. Rolling Regressions

	1. *DGP 1, R=120: horizon=12*				2. *DGP 1, R=240: horizon=12*			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.184	0.144	0.131	0.118	0.180	0.133	0.122	0.110
MSPE-normal	0.061	0.019	0.007	0.001	0.099	0.039	0.019	0.005
CH	0.108	0.078	0.052	0.023	0.134	0.091	0.066	0.038
CCS	0.201	0.169	0.167	0.168	0.205	0.175	0.158	0.153
MSPE-adj.:simul. cvs	0.238	0.195	0.175	0.148	0.228	0.185	0.173	0.155

	3. *DGP 2, R=80: horizon=4*				4. *DGP 2, R=120: horizon=4*			
	P=40	*P=80*	*P=120*	*P=160*	*P=40*	*P=80*	*P=120*	*P=160*
MSPE-adjusted	0.156	0.118	0.106	0.105	0.148	0.117	0.104	0.095
MSPE-normal	0.057	0.017	0.006	0.003	0.067	0.026	0.013	0.007
CH	0.157	0.119	0.089	0.076	0.161	0.133	0.113	0.098
CCS	0.281	0.172	0.137	0.126	0.288	0.165	0.133	0.116
MSPE-adj.:simul. cvs	0.196	0.166	0.140	0.140	0.189	0.155	0.147	0.137

B. Recursive Regressions

	1. *DGP 1, R=120: horizon=12*				2. *DGP 1, R=240: horizon=12*			
	P=120	*P=240*	*P=360*	*P=720*	*P=120*	*P=240*	*P=360*	*P=720*
MSPE-adjusted	0.177	0.136	0.115	0.091	0.176	0.133	0.111	0.092
MSPE-normal	0.077	0.039	0.021	0.008	0.103	0.055	0.037	0.017
CH	0.113	0.080	0.062	0.048	0.139	0.100	0.080	0.066
CCS	0.191	0.155	0.153	0.142	0.208	0.170	0.152	0.141
MSPE-adj.:simul. cvs	0.226	0.190	0.163	0.141	0.218	0.178	0.158	0.142

	3. *DGP 2, R=80: horizon=4*				4. *DGP 2, R=120: horizon=4*			
	P=40	*P=80*	*P=120*	*P=160*	*P=40*	*P=80*	*P=120*	*P=160*
MSPE-adjusted	0.153	0.114	0.103	0.098	0.148	0.115	0.103	0.094
MSPE-normal	0.062	0.024	0.015	0.010	0.072	0.037	0.024	0.015
CH	0.159	0.133	0.117	0.115	0.163	0.136	0.126	0.120
CCS	0.282	0.164	0.125	0.113	0.288	0.163	0.129	0.112
MSPE-adj.:simul. cvs	0.185	0.158	0.141	0.136	0.183	0.154	0.145	0.132

Notes:

1. See notes to Table 1.

2. For horizon τ, let $y_{t+\tau,t} = y_{t+\tau}+y_{t+\tau-1}+...+y_{t+1}$ be the sum of the dependent variable over the next τ periods. Using overlapping observations, predictions are made using least squares regressions of $y_{t+\tau,t}$ on X_{1t} and on X_{2t}. For given P, the number of predictions made is P-τ+1.

Table 7
Forecasts of Monthly Excess Stock Returns and Quarterly GDP Growth

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				A. Rolling Regressions					
predictand	**prediction sample**	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	**adj.**	$\hat{\sigma}_2^2$**-adj.**	**MSPE-normal**	**MSPE-adj.**	**CH**	**CCS**
(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*	*-2.13*	*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***		
				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**	*0.18*	*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) *2.07***		

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 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).

3. In column (5), "adj." is the adjustment term $P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1}-\hat{y}_{2t+1})^2$, where $\hat{y}_{1t+1}-\hat{y}_{2t+1}$ is the difference between forecasts of the two models. In column (6), "$\hat{\sigma}^2$-adj." is the difference between column (4) and column (5).

4. 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.

5. Column (9) presents the t-statistic for the Chong-Hendry statistic (3.12), column (10) the $\chi^2(1)$ (stock return) or $\chi^2(4)$ (GDP growth) statistics for the Chao et al. statistic (3.13).

6. ** 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 (2005a).

Figure 1: Null Densities of Simulated Tests, Recursive Scheme

R=240, P Varying, DGP 1

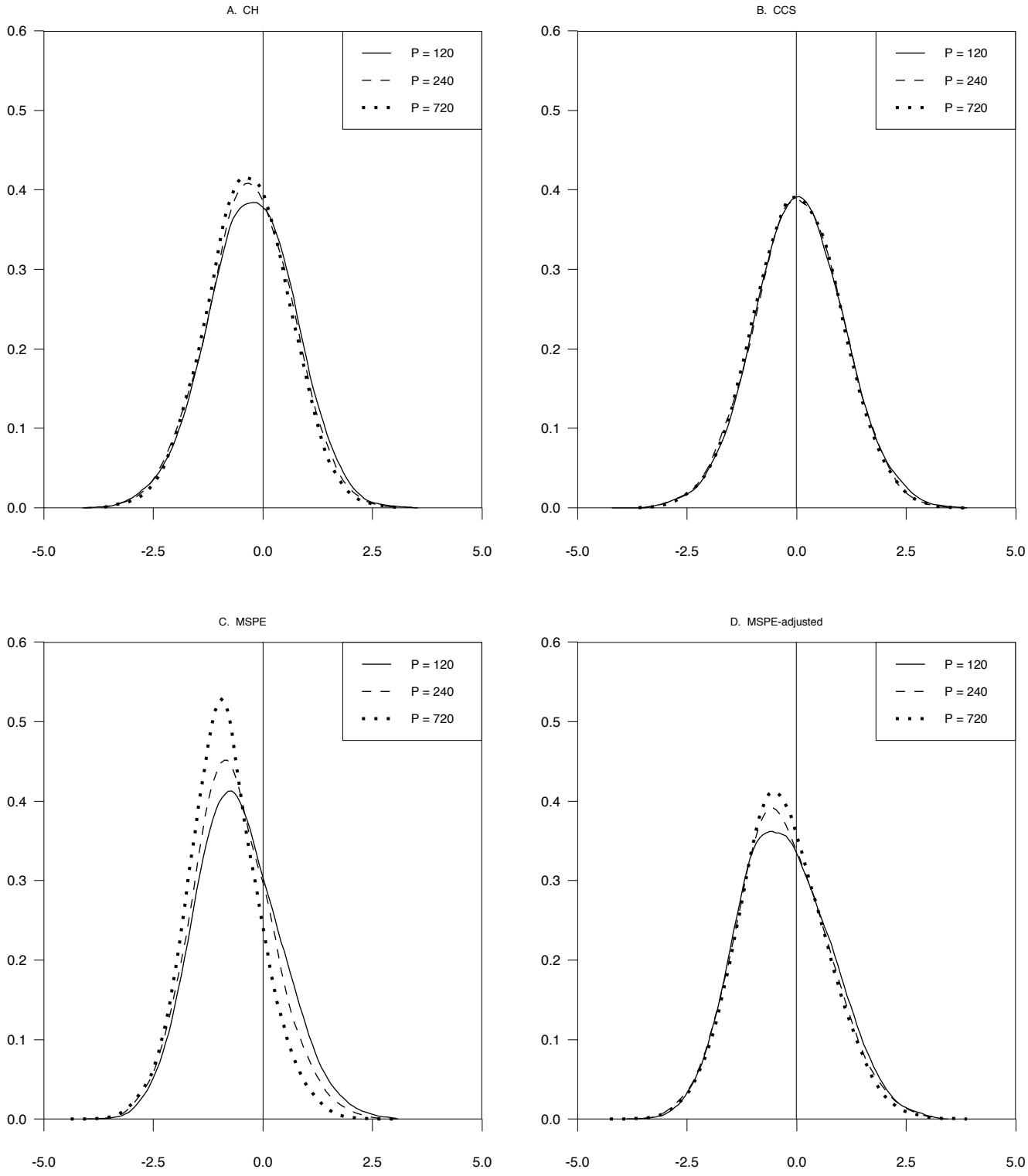


Figure 2: Null Densities of Simulated Tests, Rolling Scheme

R=240, P Varying, DGP 1