Tests for Forecast Encompassing

David I. HARVEY, Stephen J. LEYBOURNE, and Paul NEWBOLD
Department of Economics, University of Nottingham, Nottingham NG7 2RD, United Kingdom

We consider the situation in which two forecasts of the same variable are available. The possibility exists of forming a combined forecast as a weighted average of the individual ones and estimating the weights that should be optimally attached to each forecast. If the entire weight should optimally be associated with one forecast, that forecast is said to encompass the other. A natural test for forecast encompassing is based on least squares regression. We find, however, that the null distribution of this test statistic is not robust to nonnormality in the forecast errors. We discuss several alternative tests that are robust.

KEY WORDS: Combining forecasts; Conditional efficiency; Forecast evaluation; Robustness.

The evaluation of achieved forecasting records is an important aspect of practical prediction. It is frequently the case that two alternative forecasts of the same variable are available. Indeed, an analyst might create, possibly through some simple forecast-generating algorithm, a set of competitor forecasts to generate comparison with a preferred predictor. Obviously, forecasts can be compared informally or formally with actual outcomes in any number of ways. In particular, it is desirable to have available formal hypothesis-testing procedures for the analysis of competing forecasts. One possibility, analyzed extensively by Diebold and Mariano (1995), is to test the equality of forecast mean squared errors or some other measure of economic loss. As urged by Granger and Newbold (1973, 1986), however, the discovery that an analyst's preferred forecasts are better, or even "significantly better," than those from some possibly naive competitor ought not to induce complacency. A more stringent requirement would be that the competing forecasts embody no useful information absent in the preferred forecasts. Granger and Newbold proposed a formal approach to this question through the formation of composite forecasts as weighted averages of the individual forecasts. They defined the preferred forecasts as "conditionally efficient" with respect to the competing forecasts if the optimal weight attached to the latter in the composite predictor is 0. More recently, Chong and Hendry (1986) and Clements and Hendry (1993) referred to this situation as the preferred forecasts "encompassing" the competing forecasts.

Given a record of past forecast errors, it is natural to test for forecast encompassing through a simple least squares regression approach. For one-step-ahead prediction, it may be reasonable to assume, at least as a reference case, that the forecast errors are not autocorrelated so that the regression-based test is very straightforward to implement. If forecasts are for longer horizons, however, errors from optimal forecasts will be autocorrelated, for which some allowance is necessary in the development of valid tests. In Section 1 of the article, we analyze a further difficulty with the regression-based test, even for one-step prediction. It seems to us unreasonable in practice to have much faith in the assumption that forecast errors are normally distributed. Indeed, it is reasonable to suspect that forecast error distributions will often be heavy-tailed. We show in these circumstances that the regression-based test is not robust and that its application can lead to far too many rejections of a true null hypothesis of forecast encompassing. This conclusion is demonstrated analytically, and its practical force is investigated through simulations.

The problem of incorrectly sized tests introduced in Section 1 arises through conditional heteroscedasticity in the regression errors. In Section 2 of the article we discuss the application of heteroscedasticity-robust methods to alleviate the problem. The resulting test statistics are closely related to a test that follows from an application of an approach applied by Diebold and Mariano (1995) to testing the equality of expected squared forecast errors. These test procedures can be extended to yield robustness to autocorrelation as well as heteroscedasticity, as would be necessary in testing for forecast encompassing at horizons beyond one-step-ahead. Finally, Section 3 of the article reports some comparisons of the powers of these tests and a nonparametric test.

1. INCORRECT SIZE IN THE STANDARD TEST

Let (f_{1t}, f_{2t}) be two competing forecasts of the quantity y_t. In this section, we shall assume one-step-ahead prediction so that forecasts are based on information available at time $(t-1)$. It will further be assumed that the individual forecast errors have zero mean and are not autocorrelated. These two assumptions should hold approximately for reasonably well-conceived forecasts. The assumption of no autocorrelation in the forecast errors restricts our analysis to one-step-ahead prediction. In Section 2, however, we shall discuss an approach that is valid for prediction at longer horizons.

There is an extensive literature (Bates and Granger 1969; Newbold and Granger 1974; Clemen 1989; Granger 1989) on the combination of forecasts, in which the possibility is entertained of a composite forecast,

$$f_{ct} = (1-\lambda)f_{1t} + \lambda f_{2t}, \qquad 0 \le \lambda \le 1, \tag{1}$$

Journal of Business & Economic Statistics
April 1998, Vol. 16, No. 2

as a weighted average of the two individual forecasts. Then, if $e_{it} = (y_t - f_{it}), i = 1, 2$, denote the errors of the individual forecasts and ε_t is the error of the combined forecast, we can write

$$e_{1t} = \lambda(e_{1t} - e_{2t}) + \varepsilon_t. \tag{2}$$

The combined forecast will then have smaller expected squared error than f_{1t} unless the covariance between e_{1t} and $(e_{1t} - e_{2t})$ is 0. The literature on forecast combination exploits this idea, attempting to estimate λ and produce forecasts that are superior to the two individual forecasts. Granger and Newbold (1973, 1986) proposed estimation of the regression (2) to assess whether f_{2t} contains useful information not present in f_{1t}. The null hypothesis is $\lambda = 0$, and given the interpretation of the combined forecast (1), the obvious alternative is $\lambda > 0$. When the null hypothesis is true, Granger and Newbold defined f_{1t} to be "conditionally efficient" with respect to f_{2t}. An early empirical investigation along these lines was reported by Nelson (1972). Subsequently, Chong and Hendry (1986) and Clements and Hendry (1993) referred to this concept of forecast conditional efficiency as f_{1t} "encompassing" f_{2t}. Given a series of observed forecast errors $(e_{1t}, e_{2t}), t = 1, \ldots, n$, an obvious procedure is to estimate the regression (2) by ordinary least squares and apply the standard regression-based test of the null hypothesis $\lambda = 0$.

This test for forecast encompassing might be expected to perform well when the forecast errors (e_{1t}, e_{2t}) are generated by a bivariate normal distribution. It is difficult, however, to be sanguine about an assumption of normality in the forecast errors. Intuition suggests, to the contrary, that on occasions very large absolute errors might be expected so that one should be concerned about the possibility of heavy-tailed error distributions. One possibility of this sort, for example, is that the individual forecast errors obey Student's t distributions, with relatively small degrees of freedom. If the distribution of (e_{1t}, e_{2t}) is not bivariate normal, the variance of the error term conditional on the regressand in (2) may be nonconstant, and we shall allow that possibility in our subsequent analysis.

Consider the least squares estimation of (2) under the null hypothesis $\lambda = 0$, assuming for now that (e_{1t}, e_{2t}) is an independent identically distributed sequence. The regression errors are permitted to be conditionally heteroscedastic in the sense that

$$E[\varepsilon_t^2 | e_{1t} - e_{2t}] = E[e_{1t}^2 | e_{1t} - e_{2t}] = g(e_{1t} - e_{2t}). \tag{3}$$

Then, if $\hat{\lambda}$ denotes the least squares estimator of λ, under the conditions of theorem 5.3 of White (1984, p. 109),

$$D^{-1/2} n^{1/2} (\hat{\lambda} - \lambda) \xrightarrow{d} \mathrm{N}(0, 1), \tag{4}$$

where

$$\begin{aligned} D &= M^{-2} Q; \quad M = E[(e_{1t} - e_{2t})^2]; \\ Q &= \mathrm{var}\left[n^{-1/2} \sum (e_{1t} - e_{2t}) \varepsilon_t \right]. \end{aligned} \tag{5}$$

Here then we have

$$\begin{aligned} Q = E[e_{1t}^2 (e_{1t} - e_{2t})^2] &= E[(e_{1t} - e_{2t})^2 E(e_{1t}^2 | e_{1t} - e_{2t})] \\ &= E[(e_{1t} - e_{2t})^2 g(e_{1t} - e_{2t})]. \end{aligned} \tag{6}$$

The standard regression-based statistic for testing the null hypothesis is

$$R = \hat{D}^{-1/2} n^{1/2} \hat{\lambda}, \tag{7}$$

where $\hat{D} = \hat{M}^{-2} \hat{Q}, \hat{M} = n^{-1} \sum (e_{1t} - e_{2t})^2, \hat{Q} = s^2 n^{-1} \sum (e_{1t} - e_{2t})^2$, and s^2 is the residual variance from least squares estimation of (2). Clearly

$$\hat{M} \xrightarrow{p} M; \hat{Q} \xrightarrow{p} E(e_{1t}^2) E[(e_{1t} - e_{2t})^2]$$

so that

$$\hat{D} \xrightarrow{p} HD, \qquad H = \{E[(e_{1t} - e_{2t})^2 g(e_{1t} - e_{2t})]\}^{-1} \times E(e_{1t}^2) E[(e_{1t} - e_{2t})^2]. \tag{8}$$

In general, $\hat{D}$ is inconsistent for D of (4), and we have for the test statistic (7), under the null hypothesis,

$$R \xrightarrow{d} \mathrm{N}(0, H^{-1}), \tag{9}$$

where H is given by (8). Then, assuming knowledge of the elements of H, it is possible through elementary calculations to find the asymptotic size of the standard test based on the statistic (7).

To illustrate, consider the case in which the regression errors (e_{1t}, e_{2t}) are generated by the bivariate Student's t distribution of Dunnett and Sobel (1954). Let (u_{1t}, u_{2t}) be bivariate normal, with means 0, and let $\chi_{\nu,t}^2$ be an independent chi-squared random variable with ν df. Then

$$e_{it} = (\chi_{\nu,t}^2 / \nu)^{-1/2} u_{it}, \qquad i = 1, 2. \tag{10}$$

Under the null hypothesis that the forecast f_{1t} encompasses f_{2t}, we can set, without loss of generality,

$$E(u_{1t}^2) = E(u_{1t} u_{2t}) = 1; \quad E(u_{2t}^2) = \omega > 1. \tag{11}$$

This is an appealing distribution for the possible representation of forecast-errors. The individual errors e_{it} follow, up to a multiplicative constant, univariate Student's t distributions, which have heavy tails and were employed by Diebold and Mariano (1995) as a plausible possibility. Moreover, the common denominator for the e_{it} in (10) implies that, irrespective of correlation between the forecast errors, if a variable is relatively easy/difficult to predict at time t for one forecaster, the same will hold for the other.

It is straightforward to use the properties of the bivariate t distribution to calculate the quantity H of (8) and (9) and hence determine the asymptotic size of the standard regression test for forecast encompassing (e.g., see Zellner 1971, pp. 383–389). First, note that under the null hypothesis the random variables e_{1t} and $(\omega - 1)^{-1/2}(e_{1t} - e_{2t})$ are uncorrelated, follow a bivariate t_ν distribution, and have marginal t_ν distributions. Thus,

$$\begin{aligned} E(e_{1t}^2) &= (\nu - 2)^{-1} \nu \\ E[(e_{1t} - e_{2t})^2] &= (\nu - 2)^{-1} \nu (\omega - 1). \end{aligned} \tag{12}$$

Furthermore, $E[e_{1t}^2|e_{1t} - e_{2t}] = (\nu - 1)^{-1}\nu[1 + \nu^{-1}(\omega - 1)^{-1}(e_{1t} - e_{2t})^2]$, giving the function $g(e_{1t} - e_{2t})$ of (3). Then, the denominator of (8) is, for $\nu > 4$,

$$\begin{aligned} &E[(e_{1t} - e_{2t})^2 g(e_{1t} - e_{2t})] \\ &\quad = (\nu - 1)^{-1}\nu\{E[(e_{1t} - e_{2t})^2] \\ &\qquad\qquad + \nu^{-1}(\omega - 1)^{-1}E[(e_{1t} - e_{2t})^4]\} \\ &\quad = (\nu - 1)^{-1}\nu[(\nu - 2)^{-1}\nu(\omega - 1) \\ &\qquad\qquad + 3(\nu - 2)^{-1}(\nu - 4)^{-1}\nu(\omega - 1)] \\ &\quad = (\nu - 2)^{-1}(\nu - 4)^{-1}\nu^2(\omega - 1). \end{aligned} \tag{13}$$

Then, from (12) and (13), it follows that in (9) we have $H^{-1} = (\nu - 4)^{-1}(\nu - 2)$, which does not depend on ω of (11). For example, for $\nu = 6, H^{-1} = 2$, so for a nominal 5%-level test (using standard normal critical values) against a one-sided alternative, the true asymptotic size is

$$\int_{1.645}^{\infty} (4\pi)^{-1/2} \exp(-x^2/4)\, dx = .122.$$

Similarly, the asymptotic size of a nominal 10%-level test is .182. When the error-generating process is bivariate t_5, the nominal 5%-level and 10%-level tests have asymptotic sizes .171 and .230, respectively. This discussion suggests that the most commonly applied test for forecast encompassing can be seriously oversized when the joint distribution of the forecast errors is far from bivariate normal.

To assess the behavior of the standard regression-based test for forecast encompassing in finite samples, we carried out a simulation study. Forecast errors were generated, under the null hypothesis, from bivariate Student's t distributions with 5 and 6 df. Table 1 reports percentage numbers of rejections of the null hypothesis for nominal 5%-level and 10%-level tests against a one-sided alternative, in which, here and in subsequent experiments, results are based on 10,000 replications. It can be seen that the severity of the oversizing problem increases as the number of sample observations increases with very slow convergence to our asymptotic results. So, although the size of the tests is too high even for small samples, this problem is somewhat less acute than might be predicted from asymptotic theory.

2. SOME ROBUST TESTS

The null hypothesis to be tested is of zero correlation between e_{1t} and $(e_{1t} - e_{2t})$. If it can be assumed that (e_{1t}, e_{2t}) is an independent sequence, Spearman's rank correlation test, critical values for which were given, for example, by Kendall and Gibbons (1990), can be applied. Power properties for this test, which of course has correct size in finite samples, will be investigated in Section 3. In this section we explore modifications of the regression-based test that are more directly extended to the case in which the forecast errors constitute a dependent sequence.

Table 1. Empirical Sizes of Nominal 5%-Level and 10%-Level Regression-Based Tests for Forecast Encompassing

	5%-level		10%-level	
n	t_6 errors	t_5 errors	t_6 errors	t_5 errors
8	7.5	8.9	13.1	15.0
16	8.5	10.0	14.0	16.1
32	9.7	10.6	15.3	16.4
64	10.3	11.5	16.1	17.5
128	10.5	12.3	16.5	18.3
256	11.1	12.8	17.0	18.6

In Section 1, we saw, for the case in which the forecast errors were a temporally independent sequence but not normally distributed, that the standard test for forecast encompassing could be incorrectly sized due to conditional heteroscedasticity. This suggests the use of the heteroscedasticity-robust estimator of the variance of the estimated regression parameter of White (1980). More generally, this approach can be extended to dependent error sequences, as would be required for the analysis of forecasts beyond one-step-ahead, or indeed for one-step-ahead forecasts in which the analyst was unwilling to assume independence or the data failed to support this assumption.

We saw in Section 1 that difficulties can arise through the inconsistency of the standard regression-based estimator of the quantity Q of (5). Essentially the same issue is central for tests based on dependent sequences. Then, if (e_{1t}, e_{2t}) is a stationary ergodic sequence, under the conditions of theorem 5.16 of White (1984, p. 119), the result (4) continues to hold but now, modifying (5), with

$$Q = \lim_{n\to\infty} \operatorname{var}[n^{-1/2} \sum (e_{1t} - e_{2t})\varepsilon_t].$$

Estimation of Q therefore requires the estimation of $2\pi f(0)$, where $f(0)$ is the spectral density at zero frequency of $(e_{1t} - e_{2t})\varepsilon_t$. This could be achieved, plugging in estimates of ε_t, as did Newey and West (1987), or through one of the kernel and truncation lag-selection procedures discussed by Andrews (1991). Alternatively, the parametric approach of Den Haan and Levin (1994) could be employed. Here, for purposes of illustration and to aid comparison with their work, we follow the approach of Diebold and Mariano (1995). Given that optimal h-steps-ahead forecasts have errors that are at most $(h - 1)$-dependent, these authors proposed this assumption "as a reasonable benchmark," noting that it can be empirically tested. This leads to the suggestion of a rectangular kernel for spectral density estimation.

Adopting this approach, let $\hat{\lambda}$ denote the ordinary least squares estimator of λ in (2) and $\hat{\varepsilon}_t$ the residuals from the fitted regression. Then a natural estimator of Q is

$$\hat{Q}_1 = n^{-1} \sum_{\tau=-(h-1)}^{h-1} \sum_{t=|\tau|+1}^{n} (e_{1t} - e_{2t}) \times \hat{\varepsilon}_t(e_{1,t-|\tau|} - e_{2,t-|\tau|})\hat{\varepsilon}_{t-|\tau|}. \tag{14}$$

Then, following the discussion of Section 1, a modification of the standard regression test statistic (7) is provided by

$$R_1 = n^{-1/2}\hat{Q}_1^{-1/2} \sum (e_{1t} - e_{2t})^2 \hat{\lambda} = n^{1/2}\hat{Q}_1^{-1/2}\bar{d}, \tag{15}$$

where $\bar{d}$ is the sample mean of the sequence

$$d_t = (e_{1t} - e_{2t})e_{1t}. \tag{16}$$

Provided that the assumption of $(h-1)$-dependence is correct, the statistic (15) has an asymptotic standard normal distribution under the null hypothesis of forecast encompassing.

In fact, simulation evidence in the case $h = 1$ suggests that, although the actual and nominal sizes of the statistic (15) are close in large samples, they are quite far apart for small samples. This outcome follows from the fact that in this case we can write from (14)

$$\begin{aligned}\hat{Q}_1 &= n^{-1}\sum (e_{1t}-e_{2t})^2\varepsilon_t^2 - 2(\hat{\lambda}-\lambda)n^{-1} \\ &\quad \times \sum (e_{1t}-e_{2t})^3\varepsilon_t + (\hat{\lambda}-\lambda)^2 n^{-1}\sum (e_{1t}-e_{2t})^4 \\ &= n^{-1}\sum (e_{1t}-e_{2t})^2\varepsilon_t^2 - 0_P(n^{-1/2})0_P(1) \\ &\quad + 0_P(n^{-1})0_P(1).\end{aligned}$$

Thus, although $\hat{Q}_1$ is consistent for Q, convergence of the second term to 0 is likely to be slow. This being the case, an alternative possibility is to replace the estimator (14) by

$$\hat{Q}_2 = n^{-1}\sum_{\tau=-(h-1)}^{h-1}\sum_{t=|\tau|+1}^{n} d_t d_{t-|\tau|}, \tag{17}$$

yielding a test statistic R_2 of the form (15) but with $\hat{Q}_2$ in place of $\hat{Q}_1$. The estimator (17) is consistent for Q under the null hypothesis because then $\varepsilon_t = e_{1t}$, but not under the alternative, suggesting some concern about the power of the test.

Diebold and Mariano (1995) proposed a statistic for testing the equality of prediction mean squared errors based directly on the sample mean of the sequence $(e_{1t}^2 - e_{2t}^2)$.

Table 2. Empirical Sizes of Nominal 5%-Level and 10%-Level Modified Regression-Based Tests and Diebold–Mariano-type Tests for Forecast Encompassing (h = 1)

		5%-level			10%-level		
n	Test	N errors	t_6 errors	t_5 errors	N errors	t_6 errors	t_5 errors
8	R_1	10.1	12.0	13.5	15.8	18.1	19.6
	R_2	1.6	1.1	1.1	8.5	7.1	7.6
	DM	8.4	7.1	7.4	14.6	13.8	14.5
	MDM	4.4	3.3	3.4	10.2	9.0	9.5
16	R_1	8.0	9.9	11.4	13.1	15.6	17.1
	R_2	3.6	3.2	3.0	9.8	9.6	9.6
	DM	6.5	6.2	6.1	12.4	12.3	12.5
	MDM	4.9	4.3	4.3	10.5	10.2	10.4
32	R_1	6.2	8.9	9.2	11.5	14.5	14.9
	R_2	4.3	4.3	3.7	9.7	10.6	10.1
	DM	5.4	5.7	5.3	10.8	11.9	11.5
	MDM	4.8	4.8	4.3	9.9	10.9	10.4
64	R_1	6.0	7.5	7.7	11.3	13.0	13.3
	R_2	4.9	4.5	4.2	10.5	10.4	10.2
	DM	5.5	5.3	4.9	10.9	11.0	10.7
	MDM	5.1	4.8	4.5	10.7	10.5	10.3
128	R_1	5.7	6.5	6.8	10.7	12.1	12.1
	R_2	5.0	4.7	4.4	10.4	10.3	10.2
	DM	5.4	5.0	4.9	10.6	10.6	10.5
	MDM	5.2	4.8	4.6	10.4	10.4	10.3
256	R_1	5.5	5.9	6.1	10.6	11.3	11.3
	R_2	5.1	4.9	4.6	10.3	10.4	9.9
	DM	5.3	5.1	4.7	10.4	10.6	10.0
	MDM	5.2	5.0	4.6	10.4	10.4	9.9

Their approach can obviously be modified to test for forecast encompassing. Defining d_t as in (16), the null hypothesis is that $E(d_t) = 0$. The Diebold–Mariano statistic, DM, is then simply the ratio of $\bar{d}$ to its estimated standard error. Although this approach does not ostensibly rest on the prior estimation of the regression (2), it is clear that DM is identical to the statistic R_2, except that $d_t d_{t-|\tau|}$ in (17) is replaced by $(d_t - \bar{d})(d_{t-|\tau|} - \bar{d})$.

Harvey, Leybourne, and Newbold (1997) assessed the behavior of the DM test for the equality of prediction mean squared errors in moderate-sized samples and recommended two modifications. First, they proposed the modified test statistic

$$\text{MDM} = n^{-1/2}[n+1-2h+n^{-1}h(h-1)]^{1/2}\text{DM} \tag{18}$$

because this implies use of an estimator of the variance of $\bar{d}$ that is unbiased to order n^{-1}. Of course, this would continue to be so if the term $n^{-1}h(h-1)$ were omitted from (18). This term is retained on the grounds that the estimator would be exactly unbiased in the special case in which d_t is white noise. Second, by analogy with standard tests based on sample means, these authors recommended comparison of the test statistic with critical values from the t_{n-1} distribution rather than the standard normal. They provided simulation evidence demonstrating, for the problem of testing the equality of prediction mean squared errors, substantially better size properties for the MDM test than for the DM test in moderate samples. Moreover, each of the two modifications contributed appreciably to this improvement.

We ran a simulation experiment to assess finite sample sizes of the R_1, R_2, DM, and MDM tests for forecast encompassing. In keeping with the original proposal, standard normal critical values were used for the DM test, but given the results of Harvey et al. (1997), t_{n-1} critical values were used for the other three tests. Samples of independent error sequences (e_{1t}, e_{2t}) were generated, and this was taken as given, implying $h = 1$ in the test statistics, so that our results are directly comparable to those of Table 1. Errors were generated from the bivariate normal and the bivariate t distribution. Under the null hypothesis we set, without loss of generality, $\text{var}(e_{1t}) = \text{cov}(e_{1t}, e_{2t}) = 1$. Straightforward but tedious algebra then demonstrates that for all four test statistics the finite-sample null distribution is invariant to values of $\text{var}(e_{2t})$ greater than 1.

Table 2 shows the results of this simulation experiment. First, it should be noted that, as predicted by theory and by contrast with the results of Table 1, all four tests have approximately correct sizes in large samples cast-errors are generated by a bivariate t distribution. In small samples, however, the empirical sizes of these tests can deviate markedly from the nominal sizes, even when the forecast-error distribution is bivariate normal. In this respect, the MDM test performs, on the whole, somewhat better than its competitors and represents a distinct improvement on the standard regression-based test of Table 1 when the generating process is bivariate t.

Given its generally satisfactory size performance in the case of $h = 1$, we performed further simulations to investi-

gate the behavior of the MDM test for higher values of h. For comparison, the DM test was also included in this simulation experiment. Table 3 shows empirical sizes of nominal 5%-level tests when the tests were applied to forecast errors from two to eight steps ahead. In fact, we generated white-noise errors but, for h-steps-ahead forecasting, based the tests on estimated autocovariances of order up to $(h-1)$. Moreover, in these simulations the error distributions were normal. The empirical sizes for the MDM test are reassuringly close to the nominal sizes, except for long forecast horizons in small samples, in which it would be ambitious to expect great reliability. Moreover, the MDM test is generally clearly preferable to the DM test in terms of size properties, suggesting that the Harvey et al. (1997) modifications are well worth making. We checked the more general applicability of the conclusions drawn from Table 3 through some further simulation experiments, not reported in detail here. Specifically, for forecast horizons up to 4, we also generated forecast errors from bivariate t_5 and t_6 distributions. As in Table 2, the results were not dramatically different from those for the case of normal errors, confirming our conclusion that the null distribution of the test statistics is robust to heavy-tailed error distributions. Moreover, for two-steps-ahead predictions we generated series of autocorrelated forecast errors. Specifically, these errors, e_{it}^*, were generated from first-order moving average processes $e_{it}^* = e_{it} + \theta e_{i,t-1}, i = 1, 2$, where e_{it} is white noise, for $\theta = .5, .9$. The empirical sizes of the tests did not differ by much from those in the $h = 2$ column of Table 3, irrespective of whether the error distributions were bivariate normal or bivariate t. Diebold and Mariano (1995) reached a parallel conclusion when applying their approach to testing the equality of prediction mean squared errors.

Table 3. Empirical Sizes of Nominal 5%-Level Diebold–Mariano-type Tests for Forecast Encompassing: Multistep-ahead Prediction (normal errors)

n	Test	$h = 2$	$h = 4$	$h = 6$	$h = 8$
8	DM	12.6	14.3	15.5	—
	MDM	6.4	4.8	2.6	—
16	DM	10.0	11.9	12.5	12.7
	MDM	6.9	7.0	5.8	4.8
32	DM	7.3	10.0	11.0	11.6
	MDM	5.8	7.7	7.5	7.2
64	DM	6.5	8.2	9.7	10.5
	MDM	5.7	6.9	7.8	8.0
128	DM	5.6	6.5	7.4	8.5
	MDM	5.2	6.0	6.6	7.0
256	DM	5.4	5.9	6.2	6.5
	MDM	5.2	5.4	5.8	6.1

3. SOME POWER COMPARISONS

In this section, we present power comparisons of four tests for forecast encompassing in the "benchmark" case for one-step-ahead forecasts. Independent sequences of forecast errors (e_{1t}, e_{2t}) were generated from bivariate normal and bivariate t_6 distributions, setting, in the notation of Section 2, $h = 1$ in the formulas for the test statistics. To facilitate comparisons, size-adjusted powers were computed. The four tests are the standard regression-based test R, the modified regression-based test R_1, the modified Diebold–Mariano test MDM, and Spearman's rank correlation test r_s. Note that both the Diebold–Mariano test DM and the modified regression-based test R_2 have size-adjusted powers identical to that of MDM. In the former case, this is so because MDM is a constant multiple of DM, and in the latter, for $h = 1$, we have $(\text{MDM})^2 = (n - R_2^2)^{-1}(n-1)R_2^2$ so that $(\text{MDM})^2$ is a monotonic function of R_2^2.

Table 4. Estimated Size-Adjusted Powers of 5%-Level Tests for Forecast Encompassing ($h = 1$)

		High power		Moderate power	
n	Test	N errors	t_6 errors	N errors	t_6 errors
8	R	74.8	67.0	34.2	30.0
$k_1 = 1$	R_1	69.3	59.1	30.9	26.9
$k_2 = 2$	MDM	51.5	51.0	25.2	25.3
	r_s	62.1	54.4	29.3	24.2
16	R	78.3	66.1	35.4	27.9
$k_1 = 1.5$	R_1	75.3	61.8	33.0	25.7
$k_2 = 3$	MDM	68.2	59.8	29.6	26.6
	r_s	69.5	62.8	30.4	25.6
32	R	78.5	60.1	36.6	25.8
$k_1 = 2.25$	R_1	75.5	60.4	35.5	27.5
$k_2 = 4.25$	MDM	74.1	61.4	35.0	28.2
	r_s	71.6	66.3	33.2	30.0
64	R	76.7	59.8	34.7	25.4
$k_1 = 3.25$	R_1	76.3	59.2	34.9	26.0
$k_2 = 6.25$	MDM	76.0	61.5	34.2	26.8
	r_s	72.6	69.0	32.7	29.7
128	R	75.8	56.0	35.2	24.7
$k_1 = 4.75$	R_1	75.3	57.4	34.8	26.4
$k_2 = 9$	MDM	75.2	59.2	34.7	26.7
	r_s	71.9	67.9	32.8	31.0
256	R	76.2	54.0	34.1	23.6
$k_1 = 6.75$	R_1	76.2	56.2	34.5	25.2
$k_2 = 12.75$	MDM	76.2	57.6	34.5	25.8
	r_s	72.6	66.4	32.6	30.1

NOTE: k_1 is the parameter k used for high power comparisons, and k_2 is the parameter k used for moderate power comparisons.

Without loss of generality, we take $E(e_{1t}^2) = 1$. Moreover, let $E(e_{1t}e_{2t}) = \delta < 1, E(e_{1t}^2) = \omega > \delta^2$. Then, after straightforward but tedious algebra, it can be shown that, for any sample size, the powers of the tests depend only on the single parameter

$$k = (1-\delta)^{-1}(\omega - \delta^2)^{1/2}. \tag{19}$$

For each sample size, we chose, after a little experimentation, two values of the parameter k of (19)—one to achieve size-adjusted powers of around 70% and the other of around 30% for 5%-level tests against the one-sided alternative that the parameter λ in (2) is positive.

Table 4 shows the estimated size-adjusted powers for the four tests. As might be expected, the most substantial differences among the tests occur at the smallest sample sizes, where, even for forecast errors distributed as t_6, R is clearly the most powerful of the four tests. The test R_1 is somewhat less powerful than R but distinctly more powerful than MDM in small samples. By the time sample size is as large as 32, however, there is relatively little to choose among these three tests. The rank correlation test performs relatively well, particularly, as one might expect, when the error distribution is nonnormal. In large samples, it is a

little less powerful than the other three tests for normally distributed errors but noticeably more powerful when the error distribution is bivariate t_6.

Given that its nominal sizes are more reliable, the relatively poor power performance of MDM compared with R and R_1 in small samples, particularly for normally distributed errors, is disappointing. Unfortunately, as we saw in Table 2, the nominal significance levels of R_1 are unreliable in these sample sizes. Moreover, when the error distribution is nonnormal, the results of Table 1 illustrate that the nominal significance levels of R can be unreliable for any sample size. For that reason, our preference is for the MDM test over these competitors. The results of Table 2 suggest also that MDM should be preferred to R_2, with which it has identical size-adjusted power, for the same reason. We saw in Table 4 that, for samples of eight observations, in the case in which R has size-adjusted power of 74.8%, that of MDM is only 51.5%, when the forecast errors follow a bivariate normal distribution. We view this as an extreme case of the price that must be paid for using a test with reliable critical values when the error distribution is nonnormal. Of course, that price falls rapidly as the sample size increases. Finally, it is clear from Table 4 that, in the case in which the forecast errors are an independent sequence, the rank correlation test is certainly a viable alternative to MDM.

4. SUMMARY

We have analyzed the properties of several tests of the null hypothesis of forecast encompassing. This analysis has been prompted in part by lack of robustness to nonnormality in the forecast errors of the commonly applied regression-based test, which has been demonstrated both theoretically and through simulation. Four tests that do exhibit good robustness properties have been proposed and investigated. One of these was motivated by results of Diebold and Mariano (1995) on testing the equality of prediction mean squared errors. Their approach can be applied in an obvious way to testing for forecast encompassing.

We found that the Diebold–Mariano approach generates tests with good size and fairly good power properties, and we recommend that this approach be adopted. Testing for the equality of prediction mean squared errors and for forecast encompassing are companion problems, and it is certainly convenient to recommend a common general approach to the two. If that approach is to be adopted, however, we recommend the modifications introduced in Section 2—that is, comparison of the statistic (18) with critical values from the t_{n-1} distribution. As discussed by Harvey et al. (1997), that recommendation applies also to the problem of testing the equality of prediction mean squared errors. Our results suggest only one caveat to this general recommendation. For one-step-ahead forecasts, when temporal independence can be assumed and when there is a strong suspicion of heavy-tailed error distributions, rather more power can be achieved through a rank correlation test, particularly in large samples. The rank correlation test, however, is far less readily extended to the case of dependent error sequences, as would be required, for example, in tests based on forecasts at longer horizons.

ACKNOWLEDGMENTS

We are extremely grateful to a coeditor (Mark Watson) and to two anonymous referees for suggestions that greatly improved the presentation of this article.

[Received September 1996. Revised August 1997.]

REFERENCES

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Bates, J. M., and Granger, C. W. J. (1969), "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

Chong, Y. Y., and Hendry, D. F. (1986), "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies*, 53, 671–690.

Clemen, R. T. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

Clements, M. P., and Hendry, D. F. (1993), "On the Limitations of Comparing Mean Square Forecast Errors," *Journal of Forecasting*, 12, 617–637.

Den Haan, W. J., and Levin, A. (1994), "Inferences From Parametric and Nonparametric Covariance Matrix Estimation Procedures," International Finance Discussion Paper, Board of Governors of the Federal Reserve System.

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Dunnett, C. W., and Sobel, M. (1954), "A Bivariate Generalisation of Student's t-distribution, With Tables for Certain Special Cases," *Biometrika*, 41, 153–169.

Granger, C. W. J. (1989), "Combining Forecasts—Twenty Years Later," *Journal of Forecasting*, 8, 167–173.

Granger, C. W. J., and Newbold, P. (1973), "Some Comments on the Evaluation of Economic Forecasts," *Applied Economics*, 5, 35–47.

——— (1986), *Forecasting Economic Time Series* (2nd ed.), Orlando, FL: Academic Press.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291.

Kendall, M. G., and Gibbons, J. D. (1990), *Rank Correlation Methods*, London: Edward Arnold.

Nelson, C. R. (1972), "The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy," *American Economic Review*, 62, 902–917.

Newbold, P., and Granger, C. W. J. (1974), "Experience With Forecasting Univariate Time Series and the Combination of Forecasts," *Journal of the Royal Statistical Society*, Ser. A, 137, 131–165.

Newey, W. K., and West, K. D. (1987), "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

——— (1984), *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.