ELSEVIER

International Journal of Forecasting 13 (1997) 281–291

international journal of forecasting

Testing the equality of prediction mean squared errors

David Harvey, Stephen Leybourne, Paul Newbold*

Department of Economics, University of Nottingham, Nottingham, NG7 2RD, UK

Abstract

Given two sources of forecasts of the same quantity, it is possible to compare prediction records. In particular, it can be useful to test the hypothesis of equal accuracy in forecast performance. We analyse the behaviour of two possible tests, and of modifications of these tests designed to circumvent shortcomings in the original formulations. As a result of this analysis, a recommendation for one particular testing approach is made for practical applications.

Keywords: Comparing forecasts; Correlated forecast errors; Evaluation of forecasts; Non-normality

1. Introduction

It frequently happens that two (or more) competing forecasts of the same quantity are available. This occurs when an analyst wants to try alternative forecasting methodologies, when new forecasts are compared with existing forecasts, and in the analysis of the results of 'forecasting competitions'. Given observed performance records, it is inevitably the case that one set of forecasts will appear more successful than another, if only by a small amount. The question naturally arises as to how likely it is that this outcome is due to chance, i.e., viewing forecast comparison in the classical statistical hypothesis testing framework. Diebold and Mariano (1995), in an important contribution to the literature on forecast evaluation, approach forecast comparison in this framework.

Suppose that a pair of h-steps ahead forecasts have produced errors (e_{1t}, e_{2t}); $t = 1, \ldots, n$. The quality of a forecast is to be judged on some specified function $g(e)$ of the forecast error, e. Then, the null hypothesis of equality of expected forecast performance is

$$\mathrm{E}[g(e_{1t}) - g(e_{2t})] = 0 .$$

Defining

$$d_t = g(e_{1t}) - g(e_{2t}); \; t = 1, \ldots, n ,$$

it is natural to base a test on the observed sample mean:

$$\bar{d} = n^{-1} \sum_{t=1}^{n} d_t .$$

A difficulty is that the series d_t is likely to be autocorrelated. Indeed, for optimal h-steps ahead forecasts, the sequence of forecast errors follows a moving average process of order $(h-1)$. This result can be expected to hold approximately for any reasonably well-conceived set of forecasts. Therefore, in what follows it will be

* Corresponding author.

PII 0169-2070(96)00719-4

assumed for h-steps ahead forecasts that all autocorrelations of order h or higher of the sequence d_t are zero. In that case, it can be shown that the variance of $\bar{d}$ is, asymptotically,

$$V(\bar{d}) \approx n^{-1}\left[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k\right], \quad (1)$$

where γ_k is the kth autocovariance of d_t. This autocovariance can be estimated by

$$\hat{\gamma}_k = n^{-1}\sum_{t=k+1}^{n}(d_t - \bar{d})(d_{t-k} - \bar{d}). \quad (2)$$

The Diebold–Mariano test statistic is then

$$S_1 = [\hat{V}(\bar{d})]^{-1/2}\bar{d}, \quad (3)$$

where $\hat{V}(\bar{d})$ is obtained by substituting the estimates (2) in (1). Under the null hypothesis, this statistic has an asymptotic standard normal distribution, so that, in practice, tests can be very easily carried out.

Diebold and Mariano reported extensive experiments on the behaviour of this and other tests under the null hypothesis, concentrating on two-steps ahead forecasts, with mean squared error as the standard of forecast quality; that is, $g(e) = e^2$. Generally speaking, for moderately large samples, the performance was satisfactory in a wide range of situations, including contemporaneously correlated and autocorrelated forecast errors, and heavy-tailed as well as normal error distributions. However, the test was found to be quite seriously over-sized for moderate numbers of sample observations. The Diebold–Mariano test is considerably more versatile than any alternative test of equality of forecast performance, and is likely to be widely used in empirical evaluation studies. It is, therefore, worth exploring whether its behaviour in moderate-sized samples can be improved. We explore this possibility in Section 2 of the paper.

Diebold and Mariano compare their test favourably with a number of competitors, including the Morgan–Granger–Newbold test (Morgan, 1939–1940; Granger and Newbold, 1977). This particular test is considerably less versatile than the Diebold–Mariano test, but is worth consideration as it is known to have optimal properties in one particular case. Suppose we are dealing with one-step ahead forecasts, and assume that the errors are not autocorrelated. Further assume, as would often be reasonable, that the forecasts are unbiased. Then, equality of forecast mean squared errors is equivalent to equality of forecast error variances. However, it is important to allow for contemporaneous correlation between the forecast errors. Then, if the error distribution is bivariate normal, the uniformly most powerful unbiased test of the null hypothesis of equality of variances is the usual test for zero correlation between $u_{1t} = (e_{1t} - e_{2t})$ and $u_{2t} = (e_{1t} + e_{2t})$. Specifically, then, the test statistic is

$$S_2 = [(1 - r^2)/(n-1)]^{-1/2} r, \quad (4)$$

where

$$r = [\Sigma u_{1t}^2 \, \Sigma u_{2t}^2]^{-1/2} \Sigma u_{1t} u_{2t}.$$

A potential advantage of this test is that the null distribution (Student's t with $(n-1)$ degrees of freedom) of the statistic (4) is known exactly in finite samples, when the forecast errors are bivariate normal. However, Diebold and Mariano provide simulation evidence showing that the test can be seriously over-sized, even for very large samples, when the forecast errors are generated by heavy-tailed distributions. This drawback is very serious, as heavy-tailed distributions for forecast errors would seem to be quite plausible in practice. In Section 3 of the paper we analyse this phenomenon. Finally, in Section 4 of the paper we compare the power properties of different tests.

2. A modified Diebold–Mariano test

As we have noted, Diebold and Mariano presented simulation evidence suggesting that their test could be seriously over-sized in the case of two-steps ahead prediction ($h = 2$). In fact, as we shall see, this problem becomes more acute as the forecast horizon, h, increases. In this section, we explore the possibility of allevi-

ating the problem through modifications of their test procedure.

The first of our modifications relies on the use of an approximately unbiased estimator of the variance of $\bar{d}$, the sample mean of the loss differential sequence d_t. First, note that the *exact* variance is

$$V(\bar{d}) = n^{-1}\left[\gamma_0 + 2n^{-1}\sum_{k=1}^{h-1}(n-k)\gamma_k\right],$$

by contrast with expression (1), which is approximate. The estimator employed by Diebold and Mariano can then be written as

$$\hat{V}(\bar{d}) = n^{-1}\left[\hat{\gamma}_0^* + 2n^{-1}\sum_{k=1}^{h-1}(n-k)\hat{\gamma}_k^*\right], \qquad (5)$$

where

$$\hat{\gamma}_k^* = (n-k)^{-1}\sum_{t=k+1}^{n}(d_t-\bar{d})(d_{t-k}-\bar{d})$$
$$= (n-k)^{-1}n\hat{\gamma}_k .$$

Next, we require the expected values of the sample autocovariances, $\hat{\gamma}_k^*$. Assuming, without loss of generality, that d_t has mean zero, we can write

$$\sum_{t=k+1}^{n}(d_t-\bar{d})(d_{t-k}-\bar{d}) = \sum_{t=k+1}^{n}d_t d_{t-k} - (n+k)\bar{d}^2 + \bar{d}\left[\sum_{t=1}^{k}d_t + \sum_{t=n-k+1}^{n}d_t\right].$$

Taking expectations in this expression then yields:

$$E(\hat{\gamma}_k^*) = (n-k)^{-1}\left\{(n-k)\gamma_k - (n+k)V(\bar{d}) + 2n^{-1}\left[\sum_{j=1}^{k-1}(k-j)\gamma_j + \sum_{j=0}^{n-k}k\gamma_j + \sum_{j=1}^{k-1}(k-j)\gamma_{n-k+j}\right]\right\}. \qquad (6)$$

This follows since, for example

$$E\left(\bar{d}\sum_{t=1}^{k}d_t\right) = n^{-1}E\left(\sum_{t=1}^{n}d_t\sum_{t=1}^{k}d_t\right)$$
$$= n^{-1}\left[\sum_{j=1}^{k-1}(k-j)\gamma_j + \sum_{j=0}^{n-k}k\gamma_j + \sum_{j=1}^{k-1}(k-j)\gamma_{n-k+j}\right].$$

Then, assuming that k is small relative to n, the final term in (6) is of order n^{-2}, so that we have approximately

$$\mathrm{E}(\hat{\gamma}_k^*) \approx \gamma_k - (n-k)^{-1}(n+k)V(\bar{d}) \approx \gamma_k - V(\bar{d}) . \qquad (7)$$

Taking expectations in (5) and substituting (7) then yields:

$$E[\hat{V}(\bar{d})] \approx V(\bar{d}) - n^{-1}V(\bar{d})\left[1 + 2n^{-1}\sum_{k=1}^{h-1}(n-k)\right] = \frac{n+1-2h+n^{-1}h(h-1)}{n}V(\bar{d}) . \qquad (8)$$

It could be argued that the final term in the numerator of (8) can be dropped, on the ground that we have already neglected terms of this order in arriving at (7). However, we prefer to work with the approximate expression (8) on the grounds that it is exact in the special case where d_t is white noise. This is easily seen by setting $\gamma_j = 0$ $(j = 1, \ldots, n-k)$ in (6), and noting that in this particular case the variance of $\bar{d}$ is $n^{-1}\gamma_0$.

As a result of (8), it follows that employing an approximately unbiased estimator of the variance of $\bar{d}$ leads to a modified Diebold–Mariano test statistic:

$$S_1^* = \left[\frac{n+1-2h+n^{-1}h(h-1)}{n}\right]^{1/2} S_1 , \qquad (9)$$

where S_1 is the original statistic (3).

A further intuitively reasonable modification of the Diebold–Mariano test is to compare the statistic with critical values from the Student's t distribution with $(n-1)$ degrees of freedom, rather than from the standard normal distribution. Of course, such an approach would be

precisely correct in the case of one-step ahead prediction if the d_t were normally distributed. Although this normality assumption will not, in general, be correct, it nevertheless seems plausible to guess that the Student's t critical values would be more appropriate. Our modified test then involves the comparison of statistic (9) with these critical values.

We carried out a simulation study of the size properties of the original and modified Diebold–Mariano tests. Specifically, we generated independent standard normal white noise error series (e_1, e_{2t}), $t = 1, 2, \ldots, n$, for various sample sizes n. Forecasts of all horizons up to 10 were evaluated; that is, although the generated error series were white noise, this information was incorporated in the test statistics only in the case $h = 1$. We took expected squared error as the criterion of forecast quality, so that $d_t = e_{1t}^2 - e_{2t}^2$. In line with the simulation results of Diebold and Mariano, we estimated the actual size of nominal 10%-level tests against a two-sided alternative. The results are shown in Table 1. These, and all subsequent simulation results reported in this paper, are based on 10,000 replications, and were programmed in GAUSS. Besides the original and modified test, the table shows percentages of rejections for two intermediate cases, allowing an assessment of the separate impacts of our two modifications.

Diebold and Mariano reported evidence only for the case $h = 2$, where a problem of over-size was already apparent. Notice that this problem becomes increasingly severe as the forecast horizon grows. The performance of the modified test is, in all cases, considerably better. Although it is still somewhat over-sized, the performance of this test should be reasonably acceptable to practitioners, particularly if its tendency to reject a true null somewhat too often is kept in mind. Of course, the most dramatic improvements in test performance occur for the smallest sample sizes. We believe that this should be of particular interest to practitioners, as it will often be the case that very large numbers of forecasts are not available for comparison. In fact, we see in Table 1 for the longer forecast horizons that the size of the test appears at first to deteriorate with increasing n before improving again. Thus, the performance for very small sample sizes could be viewed as fortuitous.

The modified test differs in two ways from the original; statistic (9) is used rather than (3), and critical values of the Student's t rather than the standard normal distribution are employed. Each of these features contributes to the improvement in performance—the former somewhat more than the latter. To illustrate, consider the case of two-steps ahead of prediction with 16 observations. The original test rejects the null 20.3% of the time, while the comparable figure for the modified test is 14.2%. We see that comparison of S_1 with Student's t critical values gives a rejection rate of 17.8%, while comparison of S_1^* with standard normal critical values yields a rate of 16.4%.

The results reported in Table 1 are just a small subset of results we obtained from an extensive simulation study. The following conclusions also emerged from that study:

(1) *Contemporaneously correlated forecast errors*. We generated forecast errors with contemporaneous correlations of 0.5 and 0.9 and repeated our simulation experiments. The results were virtually identical to those of Table 1.

(2) *Autocorrelated forecast errors*. In developing the results of Table 1, we allowed for autocorrelation of order up to $(h-1)$ in the h-steps forecast errors, even though the errors actually generated were white noise. For two-steps ahead forecast errors, we also generated from the first-order moving average processes:

$$e_{it} = \varepsilon_{it} + \theta\varepsilon_{i,t-1}\,; \quad i = 1, 2\,,$$

where the ε_{it} are white noise, for $\theta = 0.5,\ 0.9$. Our findings differed very little from those of Table 1 for either test, irrespective of whether or not the errors were contemporaneously correlated. In fact, for both tests, the empirical sizes moved slightly closer to the nominal sizes with increasing θ.

(3) *Heavy-tailed error distributions*. We might suspect that real forecasting applications will generate occasional very large errors, so that error distributions could well have heavier tails

Table 1
Percentage of rejections of the true null hypothesis of equal prediction mean squared errors for the original and modified Diebold–Mariano test at nominal 10% level

h	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$
1	16.7	13.5	11.6	10.9	10.3	10.6	10.8
	11.0	10.8	10.3	10.4	10.0	10.5	10.7
	13.8	12.0	10.9	10.6	10.1	10.5	10.8
	8.4	9.6	9.7	10.1	9.9	10.4	10.6
2	30.0	20.3	15.1	12.4	11.5	10.9	10.5
	23.9	17.8	13.9	12.0	11.2	10.8	10.5
	21.1	16.4	13.2	11.7	11.1	10.6	10.4
	16.4	14.2	12.2	11.2	10.8	10.5	10.3
3	36.9	26.5	18.3	14.1	11.7	11.3	11.2
	30.9	24.1	17.0	13.7	11.5	11.2	11.1
	22.3	20.4	15.1	12.8	11.0	11.0	10.9
	18.1	18.5	14.3	12.2	10.7	10.8	10.9
4	43.2	30.9	21.3	15.9	13.0	11.6	11.2
	37.4	28.3	20.0	15.4	12.7	11.5	11.1
	20.9	22.1	17.2	13.9	11.8	11.1	11.0
	16.3	19.8	16.1	13.4	11.5	10.9	11.0
5	49.4	34.5	24.5	18.0	13.8	11.9	11.4
	43.5	31.8	23.4	17.4	13.5	11.7	11.3
	16.6	22.1	18.9	15.4	12.4	11.2	11.1
	12.9	19.9	17.8	14.9	12.2	11.1	11.0
6	58.4	37.3	26.7	19.6	14.8	12.2	11.8
	52.7	34.8	25.6	19.1	14.5	12.1	11.7
	13.5	21.8	19.6	16.4	13.2	11.5	11.3
	10.6	19.8	18.8	16.0	12.9	11.4	11.2
7	72.4	39.4	28.8	20.8	15.7	12.7	12.0
	68.3	36.9	27.5	20.4	15.4	12.5	11.9
	12.5	20.4	20.5	17.3	13.8	11.6	11.5
	9.9	18.2	19.5	16.8	13.6	11.6	11.4
8	–	42.6	30.8	22.9	16.3	13.1	12.0
	–	39.8	29.7	22.4	16.0	12.9	11.9
	–	19.2	21.0	18.5	14.2	12.0	11.5
	–	17.4	20.2	18.0	13.8	11.9	11.4
9	–	45.3	32.4	24.5	17.5	13.8	12.2
	–	42.7	31.3	23.9	17.2	13.8	12.2
	–	16.9	21.2	19.5	15.0	12.5	11.6
	–	15.1	20.2	19.0	14.7	12.4	11.6
10	–	49.0	33.4	25.3	17.9	14.2	12.4
	–	46.4	32.3	24.9	17.6	14.1	12.3
	–	15.5	21.1	19.6	15.4	12.7	11.8
	–	14.0	20.2	19.1	15.1	12.6	11.8

Note: The first entry in each cell is for the original test using N(0, 1) critical values, the second for the original test using t_{n-1} critical values, the third for the modified test using N(0, 1) critical values, and the fourth for the modified test using t_{n-1} critical values.

than the normal. In common with Diebold and Mariano, we generated forecast errors derived from the Student's t distribution with six degrees of freedom. Specifically, we generated (v_{1t}, v_{2t}) as independent t_6, defining the forecast errors as

$$e_{1t} = v_{1t}\,; \qquad e_{2t} = \rho v_{1t} + (1-\rho^2)^{1/2} v_{2t}\,.$$

Although, for $\rho \neq 0$, the error e_{2t} is not Student's t, we shall refer to this as a t_6 generating process. The results did not differ from those of Table 1. In fact, for the modified test, the empirical sizes tended to be closer to the nominal sizes. This phenomenon was particularly noticeable for the smallest samples sizes, and occurred also in the presence of contemporaneous correlation and autocorrelation in the errors.

In summary, then, the results reported in Table 1 would appear to be representative not only of our simulation experiments, but also of what an analyst might expect in practice. If anything, in practice, the modified test might be expected to perform even better than is indicated in the table. It seems clear that the modifications are worth making, as they add nothing to the computational burden. We believe that the modified Diebold–Mariano test should prove a valuable tool in the comparison of the quality of competing forecasts.

3. A modified Morgan–Granger–Newbold test

The Morgan–Granger–Newbold test, a variant of which has been applied, for example, by Ashley et al. (1980), is directly relevant only to the case of one-step prediction, where the forecast errors are assumed to be white noise. It is possible that the test could be extended to prediction beyond one-step ahead, though the prospects of, in effect, successfully allowing for moving average errors in a regression are problematic in moderate-sized samples. Certainly, the null distribution would no longer be known exactly in that case. Nevertheless, one-step prediction is important in its own right, so further investigation is worthwhile, given the known optimality properties of the test.

Diebold and Mariano (1995) reported some very discouraging simulation results. There is no good reason to believe that forecast errors will be normally distributed, and, indeed, intuition suggests that heavy-tailed error distributions are likely in practice. Certainly, it is desirable that any test be robust to this type of non-normality. However, Diebold and Mariano showed, for errors from a t_6 generating process, that the Morgan–Granger–Newbold test is seriously over-sized. Moreover, this problem persists, and, indeed, becomes worse as the sample size increases. Although no theoretical explanation of this phenomenon was given, such an explanation is, in fact, easily found, and provides a basis for a modification of the test that might alleviate the problem.

It is convenient to set the Morgan–Granger–Newbold test in a regression framework. Consider the regression:

$$y_t = \beta x_t + \varepsilon_t\,; \quad y_t = e_{1t} + e_{2t}\,; \quad x_t = e_{1t} - e_{2t}\,; \quad t = 1, 2, \ldots, n\,. \tag{10}$$

Then statistic (4) is identical to the usual statistic for testing $\beta = 0$; i.e.

$$S_2 = [\hat{\sigma}^2 / \Sigma x_t^2]^{-1/2} \hat{\beta}\,, \tag{11}$$

where

$$\hat{\beta} = \Sigma x_t y_t / \Sigma x_t^2\,; \qquad \hat{\sigma}^2 = (n-1)^{-1} \Sigma (y_t - \hat{\beta} x_t)^2\,,$$

and $\hat{\sigma}^2$ estimates the variance σ^2 of ε_t.

Under standard conditions, which hold here, it is known (see, for example, White, 1984, Theorem 5.3, p. 109) that

$$n^{1/2} D^{-1/2} (\hat{\beta} - \beta) \xrightarrow{\text{d}} \text{N}(0, 1)\,,$$

where

$$D = M^{-2} Q\,; \qquad Q = V(n^{-1/2} \Sigma x_t \varepsilon_t)\,;$$
$$M = E(x_t^2)\,. \tag{12}$$

The usual regression test on $\hat{\beta}$ is then based on the assumption that $E(\varepsilon_t^2 \mid x_t) = \sigma^2$, in which case it follows that $Q = \sigma^2 M$. Consequently, $D = \sigma^2 M^{-1}$, and then

$$n^{1/2}\hat{D}^{-1/2}(\hat{\beta}-\beta)\xrightarrow{\mathrm{d}}\mathrm{N}(0,1)\,, \tag{13}$$

where

$$\hat{D}=\hat{\sigma}^2(n^{-1}\Sigma x_t^2)^{-1}\,. \tag{14}$$

The difficulty in the present case is that the usual assumption that $\mathrm{E}(\varepsilon_t^2\,|\,x_t)=\sigma^2$ *does not hold*. To illustrate, it is sufficient to consider the simplest special case, where the forecast errors (e_{1t}, e_{2t}) have identical distributions, and are independent. First, we find explicit expressions for the terms in (12), under the null hypothesis $\beta=0$, so that $\varepsilon_t=(e_{1t}+e_{2t})$. We have

$$\begin{aligned}Q=V(x_t\varepsilon_t)&=E[(x_t\varepsilon_t)^2]-[E(x_t\varepsilon_t)]^2\\&=E(e_{1t}^4)+E(e_{2t}^4)-2E(e_{1t}^2e_{2t}^2)\\&\quad-[E(e_{1t}^2)-E(e_{2t}^2)]^2\,.\end{aligned}$$

Then, taking (e_{1t}, e_{2t}) to be independent identically distributed, with moments

$$\mathrm{E}(e_{it}^2)=\mu_2\,;\qquad \mathrm{E}(e_{it}^4)=\mu_4\,,$$

we have

$$Q=2(\mu_4-\mu_2^2)\,.$$

Also,

$$M=E(x_t^2)=E(e_{1t}^2+e_{2t}^2-2e_{1t}e_{2t})=2\mu_2\,,$$

under our assumptions. Then

$$D=M^{-2}Q=(K-1)/2\,,$$

where $K=\mu_4/\mu_2^2$ is the kurtosis of the assumed common distribution of the forecast errors.

Suppose, however, that D is estimated by (14). Then, under our assumptions,

$$\hat{D}\xrightarrow{p}M^{-1}V(e_{1t}+e_{2t})=(2\mu_2)^{-1}(2\mu_2)=1\,,$$

under the null hypothesis. Hence, $\hat{D}$ consistently estimates D only when the kurtosis of the error distribution is $K=3$, as for the normal. Our results demonstrate that, in the case considered here, the asymptotic distribution of the statistic on the left-hand side of (13) is normal with mean zero and variance $(K-1)/2$. For example, for the t_6 distribution, the kurtosis is 6, so that the statistic in (13) has variance 5/2. Now, assume that the null hypothesis is true, and that the forecast errors follow independent t_6 distributions. Then, asymptotically, the probability that the null hypothesis will be rejected at the 10% level against a two-sided alternative through the Morgan–Granger–Newbold test based on S_2 of (11) is

$$\int_{-\infty}^{-1.645}f(x)\,\mathrm{d}x+\int_{1.645}^{\infty}f(x)\,\mathrm{d}x=0.298\,, \tag{15}$$

obtained through numerical integration with

$$f(x)=(5\pi)^{-1/2}\,\mathrm{e}^{-x^2/5}\,.$$

It should be emphasised that this calculation applies only to the special case of *independent* t_6 error distributions. Nevertheless, the result is sufficient to allow us to predict that very severe size distortions can occur in practical applications of the Morgan–Granger–Newbold test. In fact, rather lengthier algebraic arguments can be applied to predict size distortions for the t_6 generating process with any correlation between forecast errors. The extent of those distortions depends on the correlation parameter ρ, and the larger is any positive ρ, the smaller is the size distortion. As we have seen, the root of the problem is inconsistent estimation of D in (12). This is a consequence of the fact that, although $(e_{1t}+e_{2t})$ and $(e_{1t}-e_{2t})$ are uncorrelated under the null hypothesis, *they are not, in general, independent*.

Having identified the source of the difficulty, we now attempt to modify the test, again working with the regression (10), but now employing a consistent estimator of D. Referring to (12), consistent estimation of M is straightforward through

$$\hat{M}=n^{-1}\Sigma x_t^2\,.$$

Now consider

$$Q=V(x_t\varepsilon_t)=E[(x_t\varepsilon_t)^2]-[E(x_t\varepsilon_t)]^2\,,$$

assuming, as should be reasonable for one-step prediction, that $x_t\varepsilon_t$ and $x_s\varepsilon_s$ are uncorrelated for $t\neq s$. Then, since the regression parameter in (10) is $\beta=E(x_ty_t)/E(x_t^2)$, it follows that $E(x_t\varepsilon_t)=0$, so that

$$Q = E(x_t^2 \varepsilon_t^2) .$$

Suppose that the regression (10) is estimated by ordinary least squares, yielding residuals $\hat{\varepsilon}_t$. Then a natural estimator of Q is

$$\hat{Q} = n^{-1} \Sigma x_t^2 \hat{\varepsilon}_t^2 .$$

Elementary calculations show that $\hat{Q}$ is consistent for Q, so that, in place of (14), a consistent estimator of D is given by

$$\hat{D} = \hat{M}^{-2} \hat{Q} = n \Sigma x_t^2 \hat{\varepsilon}_t^2 / (\Sigma x_t^2)^2 .$$

Therefore, in place of (11), our modified Morgan–Granger–Newbold test statistic is

$$S_2^* = [\Sigma x_t^2 \hat{\varepsilon}_t^2 / (\Sigma x_t^2)^2]^{-1/2} \hat{\beta} . \quad (16)$$

Although the null distributional result is no longer exact, a reasonable practical procedure is to compare this test statistic with critical values from the Student's t distribution with $(n-1)$ degrees of freedom.

Table 2 shows empirical sizes for the original Morgan–Granger–Newbold statistic (11) and the modified statistic (16) for nominal 10% level tests against a two-sided alternative. We considered both uncorrelated and positively contemporaneously correlated forecast errors. The errors e_{it} $(i = 1, 2)$ were taken to be identically distributed, with both the normal and t_6 generating processes employed in the simulations. Consider first the original test. Since the null distribution is known in the normal case, it is not surprising to find that, up to sampling error, the empirical sizes are equal to the nominal size in this case. However, for the t_6 error-generating process, the test is seriously over-sized. In fact, it can be seen that this problem becomes worse with increasing sample size. Recall that Eq. (15) predicts, in the case $\rho = 0$, an asymptotic size of 29.8% for a 10% level test. It appears that convergence to this limit is very slow. However, we also simulated samples of ten thousand observations, finding 29.0% rejections of the null, as a check on the validity of our asymptotic results. Those theoretical results apply only to the case $\rho = 0$. It appears from Table 2 that the problem of excess size in the test becomes less severe with increasing contemporaneous correlation between the forecast errors, which, as we have noted, can be predicted on theoretical grounds.

Table 2
Percentage of rejections of the true null hypothesis of equal one-step prediction mean squared errors for the original and modified Morgan–Granger–Newbold tests at nominal 10% level

Normal	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$
$\rho = 0$	10.2	10.3	10.0	10.1	9.7	10.4	10.7
	19.9	16.0	12.8	11.8	10.6	10.9	10.8
$\rho = 0.5$	10.0	9.9	10.3	10.3	10.2	10.6	10.4
	19.8	15.4	13.0	12.0	11.0	11.2	10.7
$\rho = 0.9$	10.1	9.9	10.2	10.4	10.0	10.5	10.1
	20.4	15.5	12.9	12.0	10.7	11.1	10.6
t_6	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$
$\rho = 0$	17.9	20.5	22.6	24.8	26.0	26.1	26.8
	26.3	21.2	18.0	15.7	12.9	12.0	10.7
$\rho = 0.5$	16.2	18.5	19.8	22.1	22.4	22.6	23.7
	25.4	20.8	17.5	14.9	12.5	11.4	10.6
$\rho = 0.9$	11.8	12.8	12.6	13.7	13.6	13.7	13.6
	22.0	18.5	15.1	13.3	12.0	10.8	10.3

Note: The first entry in each cell is for the original test, and the second for the modified test.

Turning now to the modified test, the conclusions are rather mixed. First, precisely as predicted by the theory, the modification does yield a test of the correct size for large samples. However, performance in small samples, where the modified test is over-sized, is poor. Indeed, in the smallest samples, the modified test is seriously over-sized when the error distribution is normal, and even worse than the original test for the t_6 error-generating process. We can, therefore, recommend use of the modified Morgan–Granger–Newbold test of equality of one-step prediction mean squared errors only when moderately large samples are available.

Because the test based on the modified statistic (16) performed poorly in small samples, we considered also the possibility of a non-parametric approach. The test based on (4) is simply the usual test for correlation between $(e_{1t} + e_{2t})$ and $(e_{1t} - e_{2t})$. Given our difficulties under non-normality, an obvious possibility is to employ Spearman's rank correlation test (see, for example, Kendall and Gibbons, 1990). Of course, in that case, the null distribution of the test statistic is known, and does not depend on the distribution of the forecast errors. We stress, however, that in common with the other tests of this section, this statistic is only directly applicable to one-step prediction, where the forecast errors are assumed to be white noise. The power of the rank correlation test is investigated in the next section.

4. Some power comparisons

It is, of course, the case that an analyst must be concerned with the power as well as the size of a test. However, power calculations are only relevant in circumstances where tests are correctly sized. With this point in mind, we restricted attention to the case of one-step prediction, with normally distributed error terms for the *modified* Diebold–Mariano test, the rank correlation variant of the Morgan–Granger–Newbold test and the original variant of that test. The t_6 error-generating process was also considered for the former two tests. Table 3 shows simulation results for tests against a two-sided alternative at the 10% level. The ratios, R, of the forecast error variances were selected for each sample size so that powers allowing meaningful comparisons were obtained. Also, three different values for the contemporaneous correlation ρ between the forecast errors were used.

We first note that, for the smallest sample sizes, the Morgan–Granger–Newbold test is, as might be expected on theoretical grounds, a good deal more powerful than the modified Diebold–Mariano test. However, that advantage evaporates quite rapidly with increasing sample size. (The relatively poor performance of the modified Diebold–Mariano test in the case $n = 8$ can be partly attributed to the fact that the test is somewhat under-sized for this number of sample observations). The performance of the rank correlation variant of the Morgan–Granger–Newbold test is generally roughly comparable to that of the modified Diebold–Mariano test for normal forecast errors. In particular, most of the advantage of the Morgan–Granger–Newbold test in small samples is lost when ranks are employed. However, in the case of heavy-tailed error distributions, the rank correlation test is rather more powerful than the modified Diebold–Mariano test.

We interpret the results of Table 3 as providing only a little incentive to proceed with elaborations of the Morgan–Granger–Newbold test. The original version of that test certainly has advantages in small samples, *when the error terms are normally distributed.* However, in practice it would be rash to rely on the validity of a normality assumption, and, as we have seen, the test is not robust to important departures from normality in the forecast errors. Moreover, we have seen that the modified test, based on S_2^* of (16), is only reliable in moderately large samples. But, it is clear from Table 3 that even the *original* Morgan–Granger–Newbold test has no particular advantage over the modified Diebold–Mariano test in these circumstances. The test based on rank correlations performs reasonably well in the simulations of Table 3, particularly for heavily-tailed error distributions. However, it is difficult to see how this test could

Table 3
Percentage of rejections of the false null hypothesis of equal one-step prediction mean squared errors for three tests at the 10% level

Normal	$n = 8$ $R = 3$	$n = 16$ $R = 2$	$n = 32$ $R = 1.5$	$n = 64$ $R = 1.375$	$n = 128$ $R = 1.25$	$n = 256$ $R = 1.1875$	$n = 512$ $R = 1.125$
$\rho = 0$	27.1	33.1	28.6	34.4	34.9	40.6	38.0
	28.7	33.2	27.0	32.7	32.6	37.8	35.7
	42.8	38.7	30.7	35.5	34.9	40.9	38.1
$\rho = 0.5$	32.2	39.2	33.9	41.9	42.1	48.0	46.0
	34.3	39.2	32.6	39.3	40.3	45.4	43.2
	50.9	46.4	35.8	43.2	42.7	48.5	46.1
$\rho = 0.9$	59.9	80.1	77.4	86.7	88.3	92.5	92.1
	71.0	81.3	75.0	84.3	86.1	90.3	89.6
	89.6	89.2	81.2	87.9	88.9	92.8	91.9
t_6	$n = 8$ $R = 3$	$n = 16$ $R = 2$	$n = 32$ $R = 1.5$	$n = 64$ $R = 1.375$	$n = 128$ $R = 1.25$	$n = 256$ $R = 1.1875$	$n = 512$ $R = 1.125$
$\rho = 0$	19.1	23.0	20.3	24.2	23.5	25.0	23.1
	27.6	30.9	26.5	30.9	31.0	34.4	32.7
$\rho = 0.5$	25.7	31.1	27.0	31.4	29.7	32.3	30.5
	33.6	37.9	32.4	37.9	37.2	41.5	40.5
$\rho = 0.9$	53.3	77.0	77.2	86.8	86.9	90.9	89.4
	71.6	83.5	79.0	87.2	88.2	92.7	91.6

Note: The first entry in each cell is for the modified Diebold–Mariano test, the second is for the rank correlation variant of the Morgan–Granger–Newbold test, and the third is for the original Morgan–Granger–Newbold test (normal errors only).

be extended to deal with forecasts beyond one-step ahead.

5. Summary

Testing the null hypothesis of equality of prediction mean squared errors is an important practical problem in forecast evaluation. In this paper we have examined two possible tests, and their modifications. Our recommendation is the use of the modified Diebold–Mariano test. This recommendation follows, in part, from the lack of robustness of the Morgan–Granger–Newbold test in the presence of heavy-tailed distributions of the forecast errors. Our attempts to rectify this deficiency have been only partially successful, even in the case of one-step prediction. In particular, the modified test based on S_2^* of (16) only has satisfactory size properties in moderately large samples, where the power advantages are negligible. Moreover, although the rank correlation variant of the test has some advantage in power, for heavy-tailed error distributions it is difficult to extend to predictions at longer horizons.

The modified Diebold–Mariano test has further advantages. Besides direct applicability to forecasts beyond one-step ahead, it does not rely on an assumption of forecast unbiasedness, and it can be applied to cost-of-error functions other than quadratic loss.

The modification we have applied to the Diebold–Mariano test *is* important. The original test can be quite seriously over-sized in small and moderate samples – a problem that becomes particularly acute for longer forecast horizons. Our modification goes a long way towards rectifying this problem, but does not cure it entirely. As the results of Table 1 indicate, the user of this test will have to keep in mind the issue of excess size in moderate samples. Nevertheless, we

believe that the modified Diebold–Mariano test constitutes the best available approach to assessing the significance of observed differences between the performance of two forecasts.

References

Ashley, R., C.W.J. Granger and R. Schmalensee, 1980, Advertising and aggregate consumption: An analysis of causality, *Econometrica* 48, 1149–1167.

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

Granger, C.W.J. and P. Newbold, 1977, *Forecasting Economic Time Series* (Academic Press, Orlando, FL).

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

Morgan, W.A., 1939–1940, A test for significance of the difference between the two variances in a sample from a normal bivariate population, *Biometrika* 31, 13–19.

White, H., 1984, *Asymptotic Theory for Econometricians* (Academic Press, Orlando, FL).

Biographies: David HARVEY is a research student in the Department of Economics, University of Nottingham.

Stephen LEYBOURNE is a Reader in Econometrics, University of Nottingham. He has published extensively in the area of time-series analysis.

Paul NEWBOLD is Professor of Econometrics, University of Nottingham. He has published extensively in the area of time-series analysis and forecasting.