Discussion Paper 52

Institute for Empirical Macroeconomics
Federal Reserve Bank of Minneapolis
Research Department
250 Marquette Avenue
Minneapolis, Minnesota 55480

September 1991

COMPARING PREDICTIVE ACCURACY I: AN ASYMPTOTIC TEST

Francis X. Diebold*

University of Pennsylvania

Roberto S. Mariano*

University of Pennsylvania

ABSTRACT

We propose and evaluate an explicit test of the null hypothesis of no difference in the accuracy of two competing forecasts. In contrast to previously developed tests, a wide variety of accuracy measures can be used (in particular, the loss function need not be quadratic, and need not even be symmetric), and forecast errors can be non-Gaussian, nonzero mean, serially correlated and contemporaneously correlated.

*Rob Engle, Jim Hamilton, Steve McNees, Bruce Mizrach, Marc Nerlove, Glenn Rudebusch, Dave Runkle, Ken West, and Ignacio Visco provided helpful input. We gratefully acknowledge support from the National Science Foundation, the University of Pennsylvania Research Foundation, the Institute for Empirical Macroeconomics, and the Federal Reserve Bank of Philadelphia. Ralph Bradley provided research assistance. All errors are ours.

This material is based on work supported by the National Science Foundation under Grant no. SES-8722451. The Government has certain rights to this material.

1. Introduction

Prediction is of fundamental importance in all the sciences, including economics. Forecast accuracy is of obvious importance to users of forecasts, because forecasts are used to guide decisions. Forecast accuracy is also of obvious importance to producers of forecasts, whose reputations (and fortunes) rise and fall with forecast accuracy.

Comparisons of forecast accuracy are also of importance to economists more generally, who are interested in discriminating among competing economic hypotheses (models). Predictive performance and model adequacy are inextricably linked--predictive failure implies model inadequacy.[1]

Given the obvious desirability of a formal statistical procedure for forecast accuracy comparisons, one is struck by the casual manner in which such comparisons are typically carried out. The literature contains literally hundreds of forecast accuracy comparisons; almost without exception, point estimates of forecast accuracy are examined, without questioning whether differences are statistically *significant*. Upon reflection, the reason for the casual approach is clear: correlation of forecast errors across space and time, as well as a number of additional complications, makes formal comparison of forecast accuracies

---

[1] Recall, for example, the influential work of Meese and Rogoff (1983), who argued forcefully that all members of a set of leading exchange rate models were inadequate, because their forecasts were no more accurate than those of a naive martingale model.

difficult.[2]

In this paper we propose a widely applicable test of the null hypothesis of no difference in accuracy between two forecasts. We review the small extant literature in section 2, and we propose a new and general test in section 3. Our procedure allows for a wide class of accuracy measures; this is important, because, as is well known, realistic economic loss functions frequently do not conform to stylized textbook examples like mean-squared prediction error.[3] Moreover, we allow for forecast errors that are potentially non-Gaussian, non-zero mean, serially correlated and contemporaneously correlated. In section 4 we perform a Monte Carlo experiment to evaluate the finite-sample performance of our test. Section 5 contains conclusions and directions for future research.

2. The Problem and Some Existing Approaches

It will prove useful to proceed from the most restrictive setup to the most general. Consider two forecasts, $\{\hat{y}_{it}\}_{t=1}^{T}$ and $\{\hat{y}_{jt}\}_{t=1}^{T}$, of the time series $\{y_t\}_{t=1}^{T}$. Let the associated forecast errors be $\{e_{it}\}_{t=1}^{T}$ and $\{e_{jt}\}_{t=1}^{T}$; it will prove convenient to stack the forecast errors into the (Tx1) vectors e_i and e_j. If:

---

[2] See, for example, the pessimistic assessments in Dhrymes, et al. (1972) and Howrey et al. (1974). But see also Mariano and Brown (1983, 1989, 1991).

[3] See, for example, the recent work by Leitch and Tanner (1991) and Chen and Meese (1991), who stress direction of loss, Cumby and Modest (1991), who stress market and country timing, and West, Edison and Cho (1991), who stress utility-based criteria.

(1) accuracy is measured by mean-squared prediction error (MSPE), and

(2) the forecast errors are

(2a) zero-mean

(2b) normally distributed

(2c) serially uncorrelated

(2d) contemporaneously uncorrelated,

then the null hypothesis of equal forecast accuracy corresponds to equal forecast error variances (by (1) and (2a)), and by (2b) - (2d), the ratio of sample variances has the usual F-distribution under the null hypothesis. That is, under the null hypothesis and the maintained assumptions (1), (2a) - (2d), the test statistic

$$F = \frac{\frac{e_i'e_i}{T}}{\frac{e_j'e_j}{T}} = \frac{e_i'e_i}{e_j'e_j}$$

is distributed as F(T, T).

Test statistic F is of little use in practice, however, because the conditions required to obtain its distribution are too restrictive. Let us first consider assumption (2d), contemporaneously uncorrelated forecast errors. Because the forecasts being compared are forecasts of the same economic time series, they will generally be contemporaneously correlated, producing correlation between the numerator and denominator of F, which will not then have the F distribution. This insight led Morgan (1939-1940) and Granger and Newbold (1977) to propose an

orthogonalizing transformation, which enables relaxation of assumption (2d).[4] Let $x = (e_i + e_j)$ and $z = (e_i - e_j)$. Then under the maintained assumptions (1) and (2a) - (2c), the null hypothesis of equal forecast accuracy is equivalent to zero correlation between x and z (that is, $\rho_{xz} = 0$) and the test statistic

$$MGN = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{1 - \hat{\rho}_{xz}^2}{T-1}}}$$

is distributed as Student's t with T-1 degrees of freedom, where[5]

$$\hat{\rho}_{xz} = \frac{x'z}{\sqrt{(x'x)\,(z'z)}}.$$

Let us now consider relaxing the assumptions (1), (2a) - (2c) underlying the Morgan-Granger-Newbold test. It is clear that the entire framework depends crucially on assumption (1), which cannot be relaxed. The remaining assumptions, however, can be weakened in varying degrees; we shall consider them in turn.

First, the unbiasedness assumption (2a), can be replaced by the (slightly) less restrictive assumption that both biases (b_i and b_j) are the same, while maintaining assumptions (1), (2b) and (2c). To see this, note that

---

[4] See also Young (1972).

[5] See, for example, Hogg and Craig (1978), pp. 300 - 303.

$$\gamma_{xz} = cov(x, z) = \sigma_i^2 - \sigma_j^2 + 2b_i^2 - 2b_j^2,$$

whereas the difference in MSPE's is

$$MSPE_i - MSPE_j = \sigma_i^2 - \sigma_j^2 + b_i^2 - b_j^2.$$

The two are equal if and only if $|b_i| = |b_j|$. Thus, the Morgan-Granger-Newbold assumption of zero biases can readily be replaced with one of nonzero, but equal, biases. It appears difficult, however, to allow for the more realistic possibility of nonzero and unequal biases.

Second, the normality assumption (2b) may be relaxed, while maintaining (1), (2a) and (2c), at the cost of substantial tedium involved with accounting for the higher-order moments that then enter the distribution of the sample correlation coefficient.[6]

Finally, some progress has been made at relaxing the "no serial correlation" assumption (2c), while maintaining (1), (2a) and (2b). This is important because, in general, forecast errors are serially correlated. In particular, linear least squares k-step-ahead forecast errors follow moving averages of order (k-1), as is easily seen by considering the linearly indeterministic covariance stationary process (y_t) with Wold representation

$$y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i}$$

where $b_0 = 1$, the coefficients are square summable, and the innovations are serially-uncorrelated, zero-mean, and constant-

---

[6] See, for example, Kendall and Stuart (1979), Chapter 26.

variance. The Wiener-Kolmogorov linear least squares k-step-ahead forecast is

$$\hat{y}_{t+k} = \sum_{i=0}^{\infty} b_{k+i} e_{t-i}$$

with associated prediction error

$$e_{t+k} = e_{t+k} + b_1 e_{t+k-1} + \ldots + b_{k-1} e_{t+1}$$

In practical applications, of course, the MA(k) error structure is best viewed as an approximation, because forecasts may be made from statistical models with estimated and/or time-varying parameters, and more generally, forecasts may simply be suboptimal.

The most naive (and inefficient) way to deal with the serial correlation problem, which has its roots in Tintner (1940), is simply to use only every kth observation on the forecast errors, discarding the rest. If the forecast errors really are MA(k-1), this procedure will completely purge them of serial correlation; otherwise, it may be looked upon as an approximate correction. Formally, let $(e^*_{it})_{t=1}^{[T/k]}$ and $(e^*_{jt})_{t=1}^{[T/k]}$ be series of length [T/k] consisting of every kth element of the original forecast error series $(e_{it})_{t=1}^{T}$ and $(e_{jt})_{t=1}^{T}$, where $[\cdot]$ rounds down to the nearest integer. Then, under the null hypothesis and the maintained assumptions (1), (2a) and (2b), and if the forecast errors are at most k-dependent, then the Morgan-Granger-Newbold test statistic formed from $(e^*_{it})_{t=1}^{[T/k]}$ and $(e^*_{jt})_{t=1}^{[T/k]}$ is distributed as Student's t with [T/k]-1 degrees of freedom.

More efficient procedures for dealing with serial correlation, which also allow for contemporaneous correlation, have been proposed by Meese and Rogoff (1988) and Diebold and Rudebusch (1991). Under the null hypothesis and assumptions (1), (2a) and (2b), it can be shown that[7]

$$\sqrt{T}\,\hat{\gamma}_{xz} \xrightarrow{d} N(0, \Sigma),$$

where

$$\hat{\gamma}_{xz} = \frac{x'z}{T}$$

$$\Sigma = \sum_{\tau=-\infty}^{\infty} \left[1 - \frac{|\tau|}{T}\right] \left[\gamma_{xx}(\tau)\gamma_{zz}(\tau) + \gamma_{xz}(\tau)\gamma_{zx}(\tau)\right].$$

$$\gamma_{xz}(\tau) = cov(x_t, z_{t-\tau}).$$

$$\gamma_{zx}(\tau) = cov(z_t, x_{t-\tau}).$$

$$\gamma_{xx}(\tau) = cov(x_t, x_{t-\tau}).$$

$$\gamma_{zz}(\tau) = cov(z_t, z_{t-\tau}).$$

Now, following Meese and Rogoff (1988), replace Σ with the estimate

---

[7] This is a well-known result (e.g., Priestley (1980), pp. 692-693) for the distribution of the sample cross-autocovariance function, $cov(\hat{\gamma}_{xz}(s), \hat{\gamma}_{xz}(u))$, specialized to a displacement of 0. That is, setting $s = u = 0$ gives the result stated in the text.

$$\hat{\Sigma}_a = \sum_{\tau=-S(T)}^{S(T)} \left[1 - \frac{|\tau|}{T}\right] \left[\hat{\gamma}_{xx}(\tau)\hat{\gamma}_{zz}(\tau) + \hat{\gamma}_{xz}(\tau)\hat{\gamma}_{zx}(\tau)\right].$$

where

$$\hat{\gamma}_{xz}(\tau) = \frac{1}{(T-\tau)} \sum_{t=\tau+1}^{T} x_t z_{t-\tau}$$

$$\hat{\gamma}_{zx}(\tau) = \frac{1}{(T-\tau)} \sum_{t=\tau+1}^{T} z_t x_{t-\tau}$$

$$\hat{\gamma}_{xx}(\tau) = \frac{1}{(T-\tau)} \sum_{t=\tau+1}^{T} x_t x_{t-\tau}$$

$$\hat{\gamma}_{zz}(\tau) = \frac{1}{(T-\tau)} \sum_{t=\tau+1}^{T} z_t z_{t-\tau}$$

and the truncation lag S(T) grows with the sample size, but at a slower rate.[8] This produces the test statistic,

$$MR = \frac{\hat{\gamma}_{xz}}{\sqrt{\frac{\hat{\Sigma}_a}{T}}}.$$

Under the null hypothesis and the maintained assumptions (1), (2a) and (2b), MR is asymptotically distributed as standard

---

[8] If the MA(k-1) approximation is accurate, the truncation lag need not be increased beyond k-1. The Cumby-Huisinga (1991) test provides a useful guide to the reliability of the moving-average approximation.

normal.[9]

The formula for Σ shows that the correction for serial correlation can be substantial, even if the prediction errors are only weakly serially correlated (e.g., the prediction errors might follow an MA(k-1) process with small coefficients), due to cumulation of the autocovariance terms. Conversely, if the null hypothesis and assumptions (1), (2a), (2b) and (2c) are satisfied, then all terms in Σ are zero except $\gamma_{xx}(0)$ and $\gamma_{zz}(0)$, so that MR coincides asymptotically with MGN.

Thus far we have considered relaxation of assumptions (2a) - (2c), one at a time. Simultaneous relaxation of multiple assumptions is possible within the Morgan-Granger-Newbold orthogonalizing transformation framework, but even more tedious. The distribution theory required for joint relaxation of (2b) and (2c), for example, is complicated by the presence of fourth-order cumulants in the distribution of the sample autocovariances.[10]

In the next section, however, we propose an alternative and simple testing framework that facilitates simultaneous relaxation of <u>all</u> of assumptions (2a) - (2d) <u>and assumption (1)</u>. In

---

[9] Diebold and Rudebusch (1991) make use of the closely related covariance matrix estimator

$$\hat{\Sigma}_b = \sum_{\tau=-S(T)}^{S(T)} [\hat{\gamma}_{xx}(\tau)\hat{\gamma}_{zz}(\tau) + \hat{\gamma}_{xz}(\tau)\hat{\gamma}_{zx}(\tau)] .$$

Σ_a and Σ_b are equivalent asymptotically, because the sum in Σ_a is the Caesaro sum of the sequence in Σ_b. Thus, the weighted and unweighted sums are equal, whenever the unweighted sum is convergent.

[10] See Hannan (1970), chapter 4, and Mizrach (1991).

particular, the loss function need not be quadratic (and need not even be symmetric), and forecast errors can be non-Guassian, non-zero mean, serially correlated and contemporaneously correlated. Our procedure is based upon recognition of the fact that the null of equal forecast accuracy can be mapped into the familiar null that a particular random variable has a zero mean.

3. Direct Use of Nonparametric Procedures

We define the accuracy of forecast i as (the negative of the) expected value of an arbitrary function of the forecast error, that is, $E[g(e_{it})]$. The loss function $g(\cdot)$--that is, the definition of accuracy--can be, but need not be, mean-squared prediction error or mean-absolute prediction error. The null hypothesis of equal forecast accuracy is $E[g(e_{it})] = E[g(e_{jt})]$, or $Ed_t = 0$, where $d_t \equiv [g(e_{it}) - g(e_{jt})]$. Thus, the "equal accuracy" null hypothesis is equivalent to the null hypothesis that the population mean of the time series (d_t) is 0.

Under general conditions,[11]

$$\sqrt{T}(\bar{d} - Ed_t) \xrightarrow{d} N(0, f_d(0)),$$

where

[11] See, for example, Priestley (1981) and Andrews (1991), and the references therein.

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(e_{it}) - g(e_{jt})]$$

is the sample mean loss differential,

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$$

is the spectral density of d_t at frequency zero, and $\gamma_d(\tau)$ is the autocovariance of d_t at displacement τ.[12] Thus, a natural statistic for testing the null hypothesis of equal forecast accuracy is

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}},$$

which is asymptotically N(0, 1) under the null.

The proportionality of the limiting variance to the spectral density at frequency zero is easily seen. Immediately,

$$var(\bar{d}) = \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma_d(t-s).$$

Now make the change of variables (s, t) → (s, τ), where τ = t-s, so that we sum diagonals of the variance-covariance matrix and then add the diagonal sums, rather than summing rows and adding row sums, yielding

---

[12] In words, the sample mean loss differential $\bar{d}$ is asymptotically normal and is consistent for the population expectation, but the dependence structure of the data must be taken into consideration when computing its variance.

$$var(\bar{d}) = \frac{1}{T^2} \sum_{\tau=-(T-1)}^{(T-1)} (T - |\tau|)\, \gamma_d(\tau)$$

$$= \frac{1}{T} \sum_{\tau=-(T-1)}^{(T-1)} \left(1 - \frac{|\tau|}{T}\right) \gamma_d(\tau).$$

Now recall that the spectral density function of d is defined as

$$f_d(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau)\, e^{-i\omega\tau}$$

which implies that

$$f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau).$$

Thus for large T we have

$$var(\bar{d}) \approx \frac{1}{T} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau) = \frac{2\pi}{T} f_d(0).$$

Following standard practice, we obtain a consistent estimate of $2\pi f_d(0)$ by taking a <u>weighted</u> sum of the available sample autocovariances,

$$2\pi \hat{f}_d(0) = \frac{T}{T-1} \sum_{\tau=-(T-1)}^{(T-1)} l\left(\frac{\tau}{\hat{S}(T)}\right) \hat{\gamma}_d(\tau),$$

where

$$\hat{\gamma}_d(\tau) = \begin{cases} \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} (d_t - \bar{d})(d_{t-\tau} - \bar{d}), & \text{if } \tau \geq 0 \\ \frac{1}{T-\tau} \sum_{t=-\tau+1}^{T} (d_{t+\tau} - \bar{d})(d_t - \bar{d}), & \text{if } \tau < 0, \end{cases}$$

$l(m)$ is the spectral window, and $S(T)$ is the truncation lag.

Numerous choices for $l(m)$ and $S(T)$ have been proposed. In the Monte Carlo study presented below, we use a triangular, or Bartlett, window, as in Newey and West (1987):

$$l(m) = \begin{cases} 1 - |m|, & \text{for } |m| \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

The Bartlett window guarantees positive definiteness of the estimated variance, has performed well in a variety of applications, is readily fine-tuned to exploit the special properties of forecast errors (that is, we set $S(T) = k$, the choice of which is designed to reflect the fact that k-step-ahead forecast errors are likely to be approximately characterized by moving average processes of order (k-1)), and is computationally simple enough to be amenable to our subsequent Monte Carlo analysis.[13]

4. Monte Carlo Analysis

4a. Experimental Design

---

[13] Other window shapes and bandwidth selection procedures are of course possible. Andrews (1991), for example, suggests using the quadratic spectral kernel, together with a "plug-in" automatic bandwidth selection procedure.

We evaluate the finite-sample size of test statistics F, MGN, MR and DM under the null hypothesis and various of the maintained assumptions. The design includes a variety of specifications of forecast error contemporaneous correlation, forecast error serial correlation and forecast error distributions. In order to maintain applicability of all test statistics for comparison purposes, we use quadratic loss throughout; that is, the null hypothesis is equality of MSPE's. We emphasize again, however, that an important advantage of test statistic DM in substantive economic applications--and one not shared by the others--is its direct applicability to analyses with alternative, more realistic, loss functions.

Consider first the case of Gaussian forecast errors. We draw realizations of the bivariate forecast error process, $(e_{it}, e_{jt})_{t=1}^{T}$, with varying degrees of contemporaneous and serial correlation in the generated forecast errors. This is achieved in two steps. First, we build in the desired degree of contemporaneous correlation by drawing a (2x1) forecast error innovation vector u_t from a bivariate standard normal distribution,

$$\underset{(2x1)}{u_t} = \begin{bmatrix} u_{it} \\ u_{jt} \end{bmatrix} \sim N(0_2, I_2)$$

and then premultiplying by the Choleski factor of the desired contemporaneous innovation correlation matrix. Let the desired correlation matrix be

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \rho \in [0, 1).$$

Then the Choleski factor is

$$P = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

Thus, the transformed (2x1) vector $v_t = Pu_t$ satisfies

$$\begin{bmatrix} v_{it} \\ v_{jt} \end{bmatrix} \sim N(0_2, R)$$

This operation is repeated T times, yielding $(v_{it}, v_{jt})_{t=1}^{T}$.

Second, MA(1) serial correlation (with parameter θ) is introduced by taking[14]

$$\begin{bmatrix} e_{it} \\ e_{jt} \end{bmatrix} = \begin{bmatrix} \frac{1 + \theta L}{\sqrt{1 + \theta^2}} & 0 \\ 0 & \frac{1 + \theta L}{\sqrt{1 + \theta^2}} \end{bmatrix} \begin{bmatrix} v_{it} \\ v_{jt} \end{bmatrix}, \quad t = 1, \ldots, T.$$

We consider sample sizes of T = 8, 16, 32, 64, 128, 256 and 512, contemporaneous correlation parameters of ρ = 0, .5 and .9, and moving-average parameters of θ = 0, .5, .9.

We also consider the case of non-Gaussian forecast errors. The design is the same as for the Gaussian case described above, but driven by $(u^*_{it}, u^*_{jt})'$ rather than $(u_{it}, u_{jt})'$, where

---

[14] We use $v_o = 0$. Multiplication by $(1 + \theta^2)^{-1/2}$ is done to keep the unconditional variance normalized to 1.

$$\begin{bmatrix} u_{it}^* \\ u_{jt}^* \end{bmatrix} = \begin{bmatrix} \frac{(u_{it}^2) - 1}{\sqrt{2}} \\ \frac{(u_{jt}^2) - 1}{\sqrt{2}} \end{bmatrix}.$$

Thus, the forecast error process is driven by a standardized $\chi^2(1)$ random variable.

Throughout, we perform tests at the $\alpha = .1$ level, and we perform R = 10000 Monte Carlo replications. Common random numbers are used whenever appropriate, to provide variance reduction. The truncation lag is set at 1, reflecting the fact that the experiment is designed to mimic the comparison of 2-step-ahead forecast errors, with associated MA(1) structure.

4b. Results

Results appear in tables 1 through 4 and figures 1 and 2, which show the empirical size of the various test statistics in cases of Gaussian and non-Gaussian forecast errors, as degree of contemporaneous correlation, degree of serial correlation, and sample size are varied.

Let us first discuss the case of Gaussian forecast errors.

F is correctly sized in the absence of both contemporaneous and serial correlation, but is missized in the presence of either contemporaneous or serial correlation. Serial correlation pushes empirical size above nominal size, while contemporaneous correlation pushes empirical size drastically below nominal size. In combination, and particularly for large ρ and θ,

contemporaneous correlation dominates and F is undersized.

MGN is designed to remain unaffected by contemporaneous correlation and does remain correctly sized so long as $\theta = 0$. Serial correlation, however, pushes empirical size above nominal size.

MR is drastically undersized in small samples in the presence of serial and/or contemporaneous correlation. The asymptotic distribution obtains rather quickly, however, resulting in approximately correct size for $T > 100$.

DM is only moderately oversized in small samples. In percentage terms, the small-sample size distortion of DM is much less severe than that of MR. Nominal and empirical size converge very quickly in the case of pure contemporaneous correlation, but are slower to converge when strong serial correlation is present.

To summarize the results for the Gaussian case, it is clear that the size of the MGN statistic is distorted when forecast errors are serially correlated. MR and DM, on the other hand, are approximately correctly sized for moderately large T. In very small samples, MR is undersized while DM is oversized; the size distortion associated with MR is the more severe.

Finally, let us discuss the case of non-Gaussian forecast errors. The striking result--readily apparent in figure 2--is that F, MGN and MR are drastically missized in large as well as small samples. DM, on the other hand maintains approximately correct size throughout.

5. Conclusions and Directions for Future Research

We have proposed a formal test of the null hypothesis of equal forecast accuracy. We allow the forecast errors to be non-Gaussian, non-zero mean, serially correlated and contemporaneously correlated. Finally, and importantly for applied work, we do not require the loss function to be quadratic. (That is, we do not require accuracy to be measured by mean-squared error.)

For Gaussian forecast errors, the nominal and empirical size of our test were generally the closest, although the Meese-Rogoff test was a close contender. For non-Gaussian forecast errors, however, all tests except ours were severely missized. We believe that these results, combined with the fact that our test is applicable under a wide variety of loss structures, makes our test quite attractive.

We expect that our test will be a useful addition to the applied econometrician's tool kit. We hasten to add, however, that comparison of forecast accuracy is but one of many diagnostics that should be examined when comparing models. In particular, the superiority of a particular model in terms of forecast accuracy does not necessarily imply that forecasts from other models contain no additional information. That, of course, is the well-known message of the forecast combination and encompassing literatures.[15]

---

[15] See Chong and Hendry (1986), Clemen (1989), Fair and Shiller (1990), Diebold (1989) and Ericsson (1991).

We now discuss several extensions of the results presented here, which appear to be promising directions for future research.

First, although our test performs well in samples as small as 40 or 50, fewer forecast-error observations are sometimes available in practice. Thus, it would be useful to have available an exact finite-sample test of predictive accuracy, to complement the asymptotic test presented here. Such a test could be based, for example, on either the observed forecast errors or the ranked forecast errors, using Fisher's randomization principle or the Wilcoxon's rank-sum approach, respectively. These ideas are pursued in Diebold (1991).

Second, although we have focused in this paper on the case of two forecasts of one variable, extensions in several directions would be useful:

(a) Multiple forecasts of one variable.

(b) One forecast for each of multiple variables. This would aid, for example, in assessing the relative degree of predictability of different variables in a multivariate model.

(c) Most generally, multiple forecasts for each of multiple variables.

Third, the framework developed here may be broadened to examine not only whether forecast-error loss differentials have nonzero mean, but also whether other variables may explain loss differentials. For example, one could regress the loss

differential not only on a constant, but also on a "stage of the business cycle" indicator, in order to see whether relative predictive performance differs significantly over the cycle.

Finally, the technology developed in this paper may prove useful in developing a test of exclusion restrictions in time-series regression, which is valid regardless of whether the data are stationary or integrated. The desirability of such a test is apparent from papers like Shapiro and Watson (1988), Stock and Watson (1989), Christiano and Eichenbaum (1990), Rudebusch (1990), and Toda and Phillips (1991), in which it is simultaneously apparent that:

(a) it is difficult to determine reliably the integration status of macroeconomic time series, and

(b) the conclusions of macroeconometric studies are often critically dependent on the integration status of the relevant time series.

One may proceed by noting that tests of exclusion restrictions amount to comparisons of restricted and unrestricted sums of squares. This suggests estimating the restricted and unrestricted models recursively, and then using our test of equality of the mean-squared errors of the respective one-step-ahead forecasts. Some initial progress in this direction is made by Diebold and Rudebusch (1991) in a framework of real-time causality testing.

References

Andrews, Donald (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," Econometrica, 59, 817-858.

Chen, M. and Meese, R.A. (1991), "Banking on Currency forecasts: Is Change in Money Predictable?," Manuscript, Graduate School of Business, University of California, Berkeley.

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

Christiano, L. and Eichenbaum, M. (1990), "Unit Roots in Real GNP: Do we Know, and Do we Care?," Carnegie-Rochester Conference Series on Public Policy, 32, 7-61.

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

Cumby, R.E. and Huizinga, J. (1991), "Testing the Autocorrelation Structure of Disturbances in Least Squares and Instrumental Variables Models," Manuscript, Graduate School of Business Administration, New York University, and Graduate School of Business, University of Chicago. Forthcoming, Econometrica.

Cumby, R.E. and Modest, D.M. (1987), "Testing for Market Timing Ability: A Framework for Forecast Evaluation," Journal of Financial Economics, 19, 169-189.

Diebold, F.X. (1989), "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures," International Journal of Forecasting, 5, 589-592.

Diebold, F.X. (1991), "Comparing Predictive Accuracy II: Exact Finite-Sample Tests," Manuscript in progress, Department of Economics, University of Pennsylvania.

Diebold, F.X. and Rudebusch, G.D. (1991), "Forecasting Output with the Composite Leading Index: An Ex Ante Analysis," Journal of the American Statistical Association, 86, 603-610.

Dhrymes, P.J., et al. (1972), "Criteria for Evaluation of Econometric Models," Annals of Economic and Social Measurement, 1, 291-324.

Engel, C. (1991), "Can the Markov Switching Model Forecast Exchange Rates?," Manuscript, Department of Economics, University of Washington.

Ericsson, N.R. (1991), "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration," Manuscript, International Finance Division, Federal Reserve Board.

Fair, R.C. and Shiller, R.J. (1990), "Comparing Information in Forecasts From Econometric Models," American Economic Review, 80, 375-389.

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

Hannan, E.J. (1970), Multiple Time Series. New York: John Wiley.

Hogg, R.V. and Craig, A.T. (1978), Introduction to Mathematical Statistics (Fourth Edition). New York: MacMillan.

Howrey, E.P., Klein, Lawrence R., and McCarthy, M.D. (1974), "Notes on Testing the Predictive Performance of Econometric Models," International Economic Review, 15, 366-383.

Kendall, M. and Stuart, A., (1979), The Advanced Theory of Statistics (Volume 2, Fourth Edition). New York: Oxford University Press.

Leitch, G. and Tanner, J.E. (1991), "Econometric Forecast Evaluation: Profits Versus the Conventional Error Measures," American Economic Review, 81, 580-590.

Mariano, R.S., and Brown, B.W. (1983), "Prediction-Based Tests for Misspecification in Nonlinear Simultaneous Systems," in T. Amemiya, S. Karlin and L. Goodman (eds.), Studies in Econometrics, Time Series and Multivariate Statistics, 131-151. New York: Academic Press.

Mariano, R.S., and Brown, B.W. (1989), "Stochastic Simulation, Prediction and Validation of Nonlinear Models," in L.R. Klein and J. Marquez (eds.), Economics in Theory and Practice: An Eclectic Approach, 17-36. Boston: Kluwer Academic Publishers.

Mariano, R.S., and Brown, B.W. (1989), "Stochastic Simulation Tests of Nonlinear Econometric Models," in L.R. Klein (ed.) Comparative Performance of U.S. Econometric Models, in press.

Meese, R.A. and Rogoff, K. (1983), "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?," Journal of International Economics, 14, 3-24.

Meese, R.A. and Rogoff, K. (1988), "Was it Real? The Exchange

Rate - Interest Differential Relation Over the Modern Floating-Rate Period," Journal of Finance, 43, 933-948.

Mizrach, B. (1991), "Forecast Comparison in L_2," Manuscript, Department of Finance, Wharton School, University of Pennsylvania.

Morgan, W.A. (1939-1940), "A Test for the Significance of the Difference Between the two Variances in a Sample From a Normal Bivariate Population," Biometrika, 31, 13-19.

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

Priestley, M.B. (1981), Spectral Analysis and Time Series. New York: Academic Press.

Rudebusch, Glenn D. (1990), "Trends and Random Walks in Macroeconomic Time Series: A Re-examination," Economic Activity Working Paper #105, Federal Reserve Board, Washington, DC.

Shapiro, M.D. and Watson, M.W. (1988), "Sources of Business Cycle Fluctuations," NBER Macroeconomics Annual, 111-147.

Stock, J.H. and Watson, M.W. (1989), "Interpreting the Evidence on Money-Income Causality," Journal of Econometrics, 40, 161-181.

Tintner, G. (194), The Variate-Difference Method. Bloomington: Principia Press.

Toda, H.Y. and Phillips, P.C.B. (1991), "Vector Autoregression and Causality," Cowles Foundation Discussion Paper No. 977, Department of Economics, Yale University.

West, K.D., Edison, H.J. and Cho, D. (1991), "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," Manuscript, Department of Economics, University of Wisconsin.

Young, W.E. (1971), "The Random Walk of Stock Prices: A Test of the Variance-Time Function," Econometrica, 39, 797-811.

Table 1
Empirical Size, Test Statistic F

$$F = \frac{e_i' e_i}{e_j' e_j}$$

		Gaussian			Non-Gaussian		
		θ=0.0	θ=0.5	θ=0.9	θ=0.0	θ=0.5	θ=0.9
T	ρ						
8	.0	0.106	0.132	0.147	0.224	0.212	0.212
8	.5	0.075	0.097	0.115	0.145	0.147	0.151
8	.9	0.007	0.013	0.019	0.000	0.003	0.006
16	.0	0.099	0.134	0.148	0.264	0.254	0.256
16	.5	0.074	0.102	0.118	0.185	0.187	0.190
16	.9	0.004	0.007	0.012	0.000	0.001	0.003
32	.0	0.101	0.128	0.145	0.287	0.284	0.286
32	.5	0.070	0.102	0.118	0.212	0.212	0.214
32	.9	0.003	0.006	0.010	0.000	0.001	0.002
64	.0	0.096	0.127	0.142	0.292	0.297	0.299
64	.5	0.068	0.098	0.113	0.226	0.232	0.235
64	.9	0.002	0.005	0.008	0.000	0.001	0.003
128	.0	0.101	0.125	0.140	0.300	0.307	0.309
128	.5	0.068	0.098	0.114	0.240	0.245	0.250
128	.9	0.003	0.006	0.008	0.001	0.004	0.007
256	.0	0.100	0.126	0.141	0.299	0.307	0.312
256	.5	0.068	0.098	0.112	0.237	0.248	0.252
256	.9	0.003	0.006	0.008	0.002	0.007	0.010
512	.0	0.094	0.123	0.137	0.291	0.301	0.305
512	.5	0.073	0.096	0.108	0.235	0.246	0.252
512	.9	0.002	0.005	0.008	0.004	0.011	0.015

Notes: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10000 Monte Carlo replications are performed.

Table 2
Empirical Size, Test Statistic MGN

$$MGN = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{1 - \hat{\rho}_{xz}^2}{T-1}}}$$

		Gaussian			Non-Gaussian		
		θ=0.0	θ=0.5	θ=0.9	θ=0.0	θ=0.5	θ=0.9
T	ρ						
8	.0	0.106	0.158	0.192	0.329	0.307	0.318
8	.5	0.107	0.154	0.190	0.290	0.277	0.295
8	.9	0.109	0.148	0.188	0.148	0.184	0.209
16	.0	0.103	0.157	0.189	0.420	0.396	0.409
16	.5	0.101	0.159	0.187	0.370	0.359	0.367
16	.9	0.100	0.156	0.186	0.182	0.216	0.238
32	.0	0.108	0.155	0.183	0.467	0.468	0.468
32	.5	0.100	0.159	0.190	0.410	0.413	0.419
32	.9	0.105	0.161	0.190	0.201	0.238	0.257
64	.0	0.098	0.143	0.175	0.489	0.494	0.497
64	.5	0.101	0.149	0.180	0.429	0.444	0.447
64	.9	0.098	0.148	0.175	0.226	0.260	0.278
128	.0	0.106	0.148	0.173	0.504	0.516	0.524
128	.5	0.106	0.147	0.178	0.453	0.464	0.477
128	.9	0.108	0.155	0.183	0.233	0.273	0.294
256	.0	0.104	0.151	0.178	0.507	0.525	0.532
256	.5	0.102	0.154	0.180	0.445	0.467	0.475
256	.9	0.108	0.156	0.181	0.229	0.263	0.284
512	.0	0.091	0.134	0.157	0.496	0.505	0.512
512	.5	0.108	0.151	0.171	0.435	0.455	0.466
512	.9	0.114	0.158	0.180	0.212	0.265	0.286

Notes: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10000 Monte Carlo replications are performed.

Table 3
Empirical Size, Test Statistic MR

$$MR = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{\hat{\Sigma}_a}{T}}}$$

		Gaussian			Non-Gaussian		
		θ=0.0	θ=0.5	θ=0.9	θ=0.0	θ=0.5	θ=0.9
T	ρ						
8	.0	0.003	0.002	0.001	0.054	0.011	0.005
8	.5	0.004	0.002	0.003	0.048	0.008	0.004
8	.9	0.004	0.001	0.001	0.010	0.002	0.001
16	.0	0.038	0.029	0.027	0.295	0.202	0.159
16	.5	0.037	0.027	0.026	0.252	0.168	0.128
16	.9	0.040	0.026	0.023	0.094	0.055	0.042
32	.0	0.072	0.065	0.062	0.411	0.335	0.307
32	.5	0.068	0.066	0.064	0.359	0.294	0.268
32	.9	0.070	0.064	0.060	0.161	0.130	0.118
64	.0	0.083	0.078	0.077	0.466	0.408	0.381
64	.5	0.081	0.077	0.076	0.407	0.350	0.326
64	.9	0.083	0.083	0.076	0.200	0.169	0.156
128	.0	0.096	0.088	0.089	0.491	0.447	0.426
128	.5	0.098	0.089	0.090	0.441	0.391	0.367
128	.9	0.100	0.093	0.091	0.220	0.194	0.184
256	.0	0.098	0.093	0.093	0.500	0.459	0.435
256	.5	0.098	0.098	0.097	0.439	0.397	0.378
256	.9	0.105	0.097	0.096	0.222	0.199	0.188
512	.0	0.089	0.088	0.087	0.494	0.450	0.428
512	.5	0.105	0.098	0.097	0.432	0.390	0.370
512	.9	0.112	0.101	0.104	0.209	0.197	0.187

Notes: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10000 Monte Carlo replications are performed.

Table 4
Empirical Size, Test Statistic DM

$$DM = \frac{\bar{d}}{\sqrt{\frac{\hat{f}}{T}}},$$

		Gaussian			Non-Gaussian		
		θ=0.0	θ=0.5	θ=0.9	θ=0.0	θ=0.5	θ=0.9
T	ρ						
8	.0	0.154	0.172	0.187	0.086	0.122	0.152
8	.5	0.152	0.168	0.183	0.044	0.073	0.102
8	.9	0.149	0.166	0.183	0.061	0.093	0.114
16	.0	0.126	0.151	0.167	0.069	0.090	0.117
16	.5	0.121	0.151	0.165	0.040	0.057	0.078
16	.9	0.118	0.143	0.157	0.049	0.072	0.090
32	.0	0.116	0.140	0.151	0.077	0.101	0.130
32	.5	0.109	0.141	0.154	0.059	0.077	0.098
32	.9	0.115	0.137	0.152	0.065	0.091	0.107
64	.0	0.110	0.125	0.136	0.088	0.116	0.145
64	.5	0.108	0.130	0.137	0.074	0.095	0.113
64	.9	0.104	0.126	0.137	0.078	0.107	0.123
128	.0	0.110	0.128	0.134	0.093	0.125	0.151
128	.5	0.110	0.126	0.137	0.082	0.110	0.133
128	.9	0.111	0.130	0.142	0.093	0.114	0.134
256	.0	0.106	0.125	0.135	0.099	0.134	0.154
256	.5	0.106	0.132	0.140	0.088	0.113	0.137
256	.9	0.113	0.129	0.137	0.089	0.120	0.134
512	.0	0.096	0.112	0.121	0.097	0.130	0.150
512	.5	0.112	0.124	0.131	0.087	0.123	0.142
512	.9	0.114	0.130	0.138	0.089	0.121	0.136

Notes: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10000 Monte Carlo replications are performed.

Figure 1
Empirical Test Size, Gaussian Forecast Errors

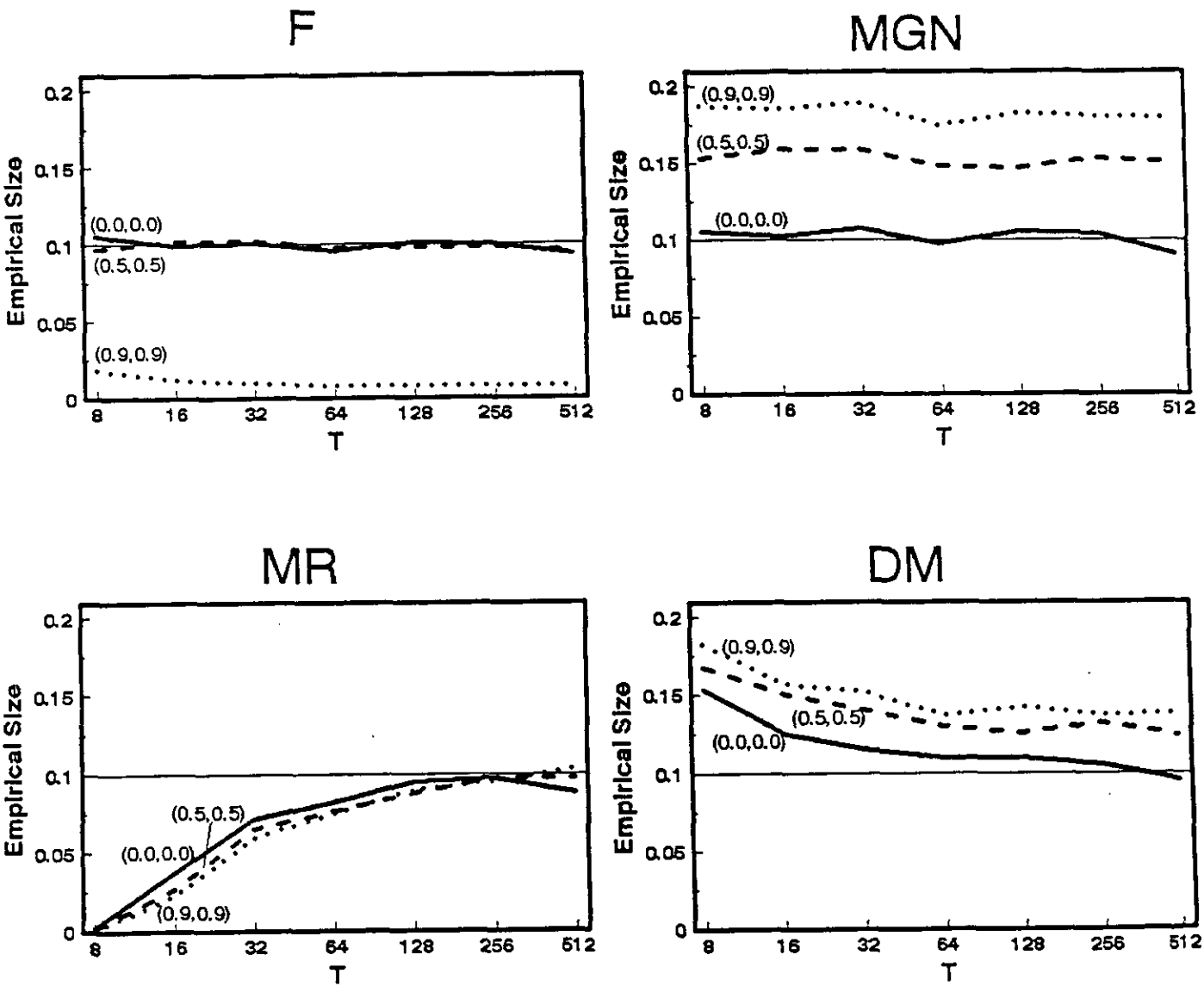

Figure 2

Empirical Test Size, Non-Gaussian Forecast Errors