G. S. Maddala and C. R. Rao, eds., *Handbook of Statistics, Vol. 14*

Forecast Evaluation and Combination*

Francis X. Diebold and Jose A. Lopez

It is obvious that forecasts are of great importance and widely used in economics and finance. Quite simply, good forecasts lead to good decisions. The importance of forecast evaluation and combination techniques follows immediately – forecast users naturally have a keen interest in monitoring and improving forecast performance. More generally, forecast evaluation figures prominently in many questions in empirical economics and finance, such as:

- Are expectations rational? (e.g., Keane and Runkle, 1990; Bonham and Cohen, 1995)
- Are financial markets efficient? (e.g., Fama, 1970, 1991)
- Do macroeconomic shocks cause agents to revise their forecasts at all horizons, or just at short- and medium-term horizons? (e.g., Campbell and Mankiw, 1987; Cochrane, 1988)
- Are observed asset returns "too volatile"? (e.g., Shiller, 1979; LeRoy and Porter, 1981)
- Are asset returns forecastable over long horizons? (e.g., Fama and French, 1988; Mark, 1995)
- Are forward exchange rates unbiased and/or accurate forecasts of future spot prices at various horizons? (e.g., Hansen and Hodrick, 1980)
- Are government budget projections systematically too optimistic, perhaps for strategic reasons? (e.g., Auerbach, 1994; Campbell and Ghysels, 1995)
- Are nominal interest rates good forecasts of future inflation? (e.g., Fama, 1975; Nelson and Schwert, 1977)

Here we provide a five-part selective account of forecast evaluation and combination methods. In the first, we discuss evaluation of a single forecast, and in particular, evaluation of whether and how it may be improved. In the second, we discuss the evaluation and comparison of the accuracy of competing forecasts. In the third, we discuss whether and how a set of forecasts may be combined to produce a superior composite forecast. In the fourth, we describe a number of

* We thank Clive Granger for useful comments, and we thank the National Science Foundation, the Sloan Foundation and the University of Pennsylvania Research Foundation for financial support.

forecast evaluation topics of particular relevance in economics and finance, including methods for evaluating direction-of-change forecasts, probability forecasts and volatility forecasts. In the fifth, we conclude.

In treating the subject of forecast evaluation, a tradeoff emerges between generality and tedium. Thus, we focus for the most part on linear least-squares forecasts of univariate covariance stationary processes, or we assume normality so that linear projections and conditional expectations coincide. We leave it to the reader to flesh out the remainder. However, in certain cases of particular interest, we do focus explicitly on nonlinearities that produce divergence between the linear projection and the conditional mean, as well as on nonstationarities that require special attention.

1. Evaluating a single forecast

The properties of optimal forecasts are well known; forecast evaluation essentially amounts to checking those properties. First, we establish some notation and recall some familiar results. Denote the covariance stationary time series of interest by y_t. Assuming that the only deterministic component is a possibly nonzero mean, μ, the Wold representation is $y_t = \mu + \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \ldots$, where $\varepsilon_t \sim WN(0, \sigma^2)$, and WN denotes serially uncorrelated (but not necessarily Gaussian, and hence not necessarily independent) white noise. We assume invertibility throughout, so that an equivalent one-sided autoregressive representation exists.

The k-step-ahead linear least-squares forecast is $\hat{y}_{t+k,t} = \mu + b_k \varepsilon_t + b_{k+1} \varepsilon_{t-1} + \ldots$, and the corresponding k-step-ahead forecast error is

$$e_{t+k,t} = y_{t+k} - \hat{y}_{t+k,t} = \varepsilon_{t+k} + b_1 \varepsilon_{t+k-1} + \ldots + b_{k-1} \varepsilon_{t+1} \quad . \tag{1}$$

Finally, the k-step-ahead forecast error variance is

$$\sigma_k^2 = \mathrm{var}\left(e_{t+k,t}\right) = \sigma^2 \left(\sum_{i=1}^{k-1} b_i^2 \right) \quad . \tag{2}$$

Four key properties of errors from optimal forecasts, which we discuss in greater detail below, follow immediately:

(1) Optimal forecast errors have a zero mean (follows from (1));
(2) 1-step-ahead optimal forecast errors are white noise (special case of (1) corresponding to $k = 1$);
(3) k-step-ahead optimal forecast errors are at most MA(k–1) (general case of (1));
(4) The k-step-ahead optimal forecast error variance is non-decreasing in k (follows from (2)).

Before proceeding, we now describe some exact distribution-free nonparametric tests for whether an independently (but not necessarily identically) distributed series has a zero median. The tests are useful in evaluating the properties of optimal forecast errors listed above, as well as other hypotheses that will concern us later. Many such tests exist; two of the most popular, which we use repeatedly, are the sign test and the Wilcoxon signed-rank test.

Denote the series being examined by x_t, and assume that T observations are available. The sign test proceeds under the null hypothesis that the observed series is independent with a zero median.[1] The intuition and construction of the test statistic are straightforward – under the null, the number of positive observations in a sample of size T has the binomial distribution with parameters T and 1/2. The test statistic is therefore simply

$$S = \sum_{t=1}^{T} I_+(x_t) \ ,$$

where

$$I_+(x_t) = \begin{cases} 1 & \text{if } x_t > 0 \ , \\ 0 & \text{otherwise.} \end{cases}$$

In large samples, the studentized version of the statistic is standard normal,

$$\frac{S - T/2}{\sqrt{T/4}} \overset{a}{\sim} N(0,1) \ .$$

Thus, significance may be assessed using standard tables of the binomial or normal distributions.

Note that the sign test does not require distributional symmetry. The Wilcoxon signed-rank test, a related distribution-free procedure, *does* require distributional symmetry, but it can be more powerful than the sign test in that case. Apart from the additional assumption of symmetry, the null hypothesis is the same, and the test statistic is the sum of the ranks of the absolute values of the positive observations,

$$W = \sum_{t=1}^{T} I_+(x_t)\text{Rank}(|x_t|) \ ,$$

where the ranking is in increasing order (e.g., the largest absolute observation is assigned a rank of T, and so on). The intuition of the test is simple – if the underlying distribution is symmetric about zero, a "very large" (or "very small") sum of the ranks of the absolute values of the positive observations is "very unlikely." The exact finite-sample null distribution of the signed-rank statistic is free from nuisance parameters and invariant to the true underlying distribution, and it has been tabulated. Moreover, in large samples, the studentized version of the statistic is standard normal,

[1] If the series is symmetrically distributed, then a zero median of course corresponds to a zero mean.

$$\frac{W - [T(T+1)]/4}{\sqrt{[T(T+1)(2T+1)]/24}} \overset{a}{\sim} N(0,1) \ .$$

Testing properties of optimal forecasts

Given a track record of forecasts, $\hat{y}_{t+k,t}$, and corresponding realizations, y_{t+k}, forecast users will naturally want to assess forecast performance. The properties of optimal forecasts, cataloged above, can readily be checked.

a. Optimal forecast errors have a zero mean

A variety of standard tests of this hypothesis can be performed, depending on the assumptions one is willing to maintain. For example, if $e_{t+k,t}$ is Gaussian white noise (as might be the case for 1-step-ahead errors), then the standard t-test is the obvious choice because it is exact and uniformly most powerful. If the errors are non-Gaussian but remain independent and identically distributed (iid), then the t-test is still useful asymptotically. However, if more complicated dependence or heterogeneity structures are (or may be) operative, then alternative tests are required, such as those based on the generalized method of moments.

It would be unfortunate if non-normality or richer dependence/heterogeneity structures mandated the use of asymptotic tests, because sometimes only short track records are available. Such is not the case, however, because exact distribution-free nonparametric tests are often applicable, as pointed out by Campbell and Ghysels (1995). Although the distribution-free tests do require independence (sign test) and independence and symmetry (signed-rank test), they do not require normality or identical distributions over time. Thus, the tests are automatically robust to a variety of forecast error distributions, and to heteroskedasticity of the independent but not identically distributed type.

For $k > 1$, however, even optimal forecast errors are likely to display serial correlation, so the nonparametric tests must be modified. Under the assumption that the forecast errors are $(k-1)$-dependent, each of the following k series of forecast errors will be free of serial correlation: $\{e_{1+k,1}, e_{1+2k,1+k}, e_{1+3k,1+2k}, \ldots\}$, $\{e_{2+k,2}, e_{2+2k,2+k}, e_{2+3k,2+2k}, \ldots\}$, $\{e_{3+k,3}, e_{3+2k,3+k}, e_{3+3k,3+2k}, \ldots\}, \ldots, \{e_{2k,k}, e_{3k,2k}, e_{4k,3k}, \ldots\}$. Thus, a Bonferroni bounds test (with size bounded above by α) is obtained by performing k tests, each of size α/k, on each of the k error series, and rejecting the null hypothesis if the null is rejected for *any* of the series. This procedure is conservative, even asymptotically. Alternatively, one could use just one of the k error series and perform an exact test at level α, at the cost of reduced power due to the discarded observations.

In concluding this section, let us stress that the nonparametric distribution-free tests are neither unambiguously "better" nor "worse" than the more common tests; rather, they are useful in different situations and are therefore complementary. To their credit, they are often exact finite-sample tests with good finite-sample power, and they are insensitive to deviations from the standard assumptions of normality and homoskedasticity required to justify more standard tests in small samples. Against them, however, is the fact that they require independence of the forecast errors, an assumption even stronger than conditional-mean independence, let alone linear-projection independence. Furthermore, although the nonparametric tests can be modified to allow for k-dependence, a possibly substantial price must be paid either in terms of inexact size or reduced power.

b. 1-Step-ahead optimal forecast errors are white noise

More precisely, the errors from linear least squares forecasts are linear-projection independent, and the errors from least squares forecasts are conditional-mean independent. The errors never need be fully serially independent, because dependence can always enter through higher moments, as for example with the conditional-variance dependence of GARCH processes.

Under various sets of maintained assumptions, standard asymptotic tests may be used to test the white noise hypothesis. For example, the sample autocorrelation and partial autocorrelation functions, together with Bartlett asymptotic standard errors, may be useful graphical diagnostics in that regard. Standard tests based on the serial correlation coefficient, as well as the Box-Pierce and related statistics, may be useful as well.

Dufour (1981) presents adaptations of the sign and Wilcoxon signed-rank tests that yield exact tests for serial dependence in 1-step-ahead forecast errors, without requiring normality or identical forecast error distributions. Consider, for example, the null hypothesis that the forecast errors are independent and symmetrically distributed with zero median. Then median $(e_{t+1,t}e_{t+2,t+1}) = 0$; that is, the product of two symmetric independent random variables with zero median is itself symmetric with zero median. Under the alternative of positive serial dependence, median $(e_{t+1,t}e_{t+2,t+1}) > 0$, and under the alternative of negative serial dependence, median $(e_{t+1,t}e_{t+2,t+1}) < 0$. This suggests examining the cross-product series $z_t = e_{t+1,t}e_{t+2,t+1}$ for symmetry about zero, the obvious test for which is the signed-rank test, $W_D = \sum_{t=1}^{T} I_+(z_t)\text{Rank}(|z_t|)$. Note that the z_t sequence will be serially dependent even if the $e_{t+1,t}$ sequence is not, in apparent violation of the conditions required for validity of the signed-rank test (applied to z_t). Hence the importance of Dufour's contribution – Dufour shows that the serial correlation is of no consequence and that the distribution of W_D is the same as that of W.

c. k-Step-ahead optimal forecast errors are at most MA(k–1)

Cumby and Huizinga (1992) develop a useful asymptotic test for serial dependence of order greater than $k - 1$. The null hypothesis is that the $e_{t+k,t}$ series is MA(q) ($0 \leq q \leq k - 1$) against the alternative hypothesis that at least one autocorrelation is nonzero at a lag greater than $k - 1$. Under the null, the sample autocorrelations of $e_{t+k,t}, \hat{\rho} = [\hat{\rho}_{q+1}, \ldots, \hat{\rho}_{q+s}]$, are asymptotically distributed $\sqrt{T}\hat{\rho} \sim N(0, V)$.[2] Thus,

[2] s is a cutoff lag selected by the user.

$$C = T\hat{\rho}'\hat{V}^{-1}\hat{\rho}$$

is asymptotically distributed as χ^2_s under the null, where $\hat{V}$ is a consistent estimator of V.

Dufour's (1981) distribution-free nonparametric tests may also be adapted to provide a finite-sample bounds test for serial dependence of order greater than $k-1$. As before, separate the forecast errors into k series, each of which is serially independent under the null of $(k-1)$-dependence. Then, for each series, take $z_{k,t} = e_{t+k,t}e_{t+2k,t+k}$ and reject at significance level bounded above by α if one or more of the subset test statistics rejects at the α/k level.

d. The k-step-ahead optimal forecast error variance is non-decreasing in k
The k-step-ahead forecast error variance, $\sigma^2_k = \text{var}(e_{t+k,t}) = \sigma^2(\sum_{i=1}^{k-1} b_i^2)$, is non-decreasing in k. Thus, it is often useful simply to examine the sample k-step-ahead forecast error variances as a function of k, both to be sure the condition appears satisfied and to see the pattern with which the forecast error variance grows with k, which often conveys useful information.[3] Formal inference may also be done, so long as one takes care to allow for dependence of the sample variances across horizons.

Assessing optimality with respect to an information set

The key property of optimal forecast errors, from which all others follow (including those cataloged above), is unforecastability on the basis of information available at the time the forecast was made. This is true regardless of whether linear-projection optimality or conditional-mean optimality is of interest, regardless of whether the relevant loss function is quadratic, and regardless of whether the series being forecast is stationary.

Following Brown and Maital (1981), it is useful to distinguish between partial and full optimality. Partial optimality refers to unforecastability of forecast errors with respect to some subset, as opposed to all subsets, of available information, Ω_t. Partial optimality, for example, characterizes a situation in which a forecast is optimal with respect to the information used to construct it, but the information used was not all that *could* have been used. Thus, each of a set of competing forecasts may have the partial optimality property if each is optimal with respect to its own information set.

One may test partial optimality via regressions of the form $e_{t+k,t} = \alpha' x_t + u_t$, where $x_t \subset \Omega_t$. The particular case of testing partial optimality with respect to $\hat{y}_{t+k,t}$ has received a good deal of attention, as in Mincer and Zarnowitz (1969). The relevant regression is $e_{t+k,t} = \alpha_0 + \alpha_1 \hat{y}_{t+k,t} + u_t$ or $y_{t+k} = \beta_0 + \beta_1 \hat{y}_{t+k,t} + u_t$, where partial optimality corresponds to $(\alpha_0, \alpha_1) = (0, 0)$ or $(\beta_0, \beta_1) = (0, 1)$.[4] One

[3] Extensions of this idea to nonstationary long-memory environments are developed in Diebold and Lindner (1995).

[4] In such regressions, the disturbance should be white noise for 1-step-ahead forecasts but may be serially correlated for multi-step-ahead forecasts.

may also expand the regression to allow for various sorts of nonlinearity. For example, following Ramsey (1969), one may test whether all coefficients in the regression $e_{t+k,t} = \sum_{j=0}^{J} \alpha_j \hat{y}_{t+k,t}^{j} + u_t$ are zero.

Full optimality, in contrast, requires the forecast error to be unforecastable on the basis of *all* information available when the forecast was made (that is, the entirety of Ω_t). Conceptually, one could test full rationality via regressions of the form $e_{t+k,t} = \alpha' x_t + u_t$. If $\alpha = 0$ for all $x_t \subseteq \Omega_t$, then the forecast is fully optimal. In practice, one can never test for full optimality, but rather only partial optimality with respect to increasing information sets.

Distribution-free nonparametric methods may also be used to test optimality with respect to various information sets. The sign and signed-rank tests, for example, are readily adapted to test orthogonality between forecast errors and available information, as proposed by Campbell and Dufour (1991, 1995). If, for example, $e_{t+1,t}$ is linear-projection independent of $x_t \in \Omega_t$, then $\text{cov}(e_{t+1,t}, x_t) = 0$. Thus, in the symmetric case, one may use the signed-rank test for whether $\text{E}[z_t] = \text{E}[e_{t+1,t} x_t] = 0$, and more generally, one may use the sign test for whether $\text{median}(z_t) = \text{median}(e_{t+1,t} x_t) = 0$.[5] The relevant sign and signed-rank statistics are $S_\perp = \sum_{t=1}^{T} I_+(z_t)$ and $W_\perp = \sum_{t=1}^{T} I_+(z_t)\text{Rank}(|z_t|)$. Moreover, one may allow for nonlinear transformations of the elements of the information set, which is useful for assessing conditional-mean as opposed to simply linear-projection independence, by taking $z_t = e_{t+1,t} g(x_t)$, where $g(.)$ is a nonlinear function of interest. Finally, the tests can be generalized to allow for k-step-ahead forecast errors as before. Simply take $z_t = e_{t+k,t} g(x_t)$, divide the z_t series into the usual k subsets, and reject the orthogonality null at significance level bounded by α if any of the subset test statistics are significant at the α/k level.[6]

2. Comparing the accuracy of multiple forecasts

Measures of forecast accuracy

In practice, it is unlikely that one will ever stumble upon a fully-optimal forecast; instead, situations often arise in which a number of forecasts (all of them suboptimal) are compared and possibly combined. The crucial object in measuring forecast accuracy is the loss function, $L(y_{t+k}, \hat{y}_{t+k,t})$, often restricted to $L(e_{t+k,t})$, which charts the "loss," "cost" or "disutility" associated with various pairs of forecasts and realizations. In addition to the shape of the loss function, the forecast horizon (k) is also of crucial importance. Rankings of forecast accuracy

[5] Again, it is not obvious that the conditions required for application of the sign or signed-rank test to z_t are satisfied, but they are; see Campbell and Dufour (1995) for details.

[6] Our discussion has implicitly assumed that both $e_{t+1,t}$ and $g(x_t)$ are centered at zero. This will hold for $e_{t+1,t}$ if the forecast is unbiased, but there is no reason why it should hold for $g(x_t)$. Thus, in general, the test is based on $g(x_t) - \mu_t$, where μ_t is a centering parameter such as the mean, median or trend of $g(x_t)$. See Campbell and Dufour (1995) for details.

may be very different across different loss functions and/or different horizons. This result has led some to argue the virtues of various "universally applicable" accuracy measures. Clements and Hendry (1993), for example, argue for an accuracy measure under which forecast rankings are invariant to certain transformations.

Ultimately, however, the appropriate loss function depends on the situation at hand. As stressed by Diebold (1993) among many others, forecasts are usually constructed for use in particular decision environments; for example, policy decisions by government officials or trading decisions by market participants. Thus, the appropriate accuracy measure arises from the loss function faced by the forecast user. Economists, for example, may be interested in the profit streams (e.g., Leitch and Tanner, 1991, 1995; Engle et al., 1993) or utility streams (e.g., McCulloch and Rossi, 1990; West, Edison and Cho, 1993) flowing from various forecasts.

Nevertheless, let us discuss a few stylized statistical loss functions, because they are used widely and serve as popular benchmarks. Accuracy measures are usually defined on the forecast errors, $e_{t+k,t} = y_{t+k} - \hat{y}_{t+k,t}$, or percent errors, $p_{t+k,t} = (y_{t+k} - \hat{y}_{t+k,t})/y_{t+k}$. For example, the mean error, $\mathrm{ME} = \frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}$, and mean percent error, $\mathrm{MPE} = \frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}$, provide measures of bias, which is one component of accuracy.

The most common overall accuracy measure, by far, is mean squared error, $\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}^2$, or mean squared percent error, $\mathrm{MSPE} = \frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}^2$. Often the square roots of these measures are used to preserve units, yielding the root mean squared error, $\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}^2}$, and the root mean squared percent error, $\mathrm{RMSPE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}^2}$. Somewhat less popular, but nevertheless common, accuracy measures are mean absolute error, $\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} |e_{t+k,t}|$, and mean absolute percent error, $\mathrm{MAPE} = \frac{1}{T}\sum_{t=1}^{T} |p_{t+k,t}|$.

MSE admits an informative decomposition into the sum of the variance of the forecast error and its squared bias,

$$\begin{aligned}\mathrm{MSE} = \mathrm{E}\left[\left(y_{t+k} - \hat{y}_{t+k,t}\right)^2\right] &= \mathrm{var}\left(y_{t+k} - \hat{y}_{t+k,t}\right) \\ &+ \left(\mathrm{E}[y_{t+k}] - \mathrm{E}\left[\hat{y}_{t+k,t}\right]\right)^2 ,\end{aligned}$$

or equivalently

$$\begin{aligned}\mathrm{MSE} = \mathrm{var}(y_{t+k}) + \mathrm{var}\left(\hat{y}_{t+k,t}\right) &- 2\,\mathrm{cov}\left(y_{t+k}, \hat{y}_{t+k,t}\right) \\ &+ \left(\mathrm{E}[y_{t+k}] - \mathrm{E}\left[\hat{y}_{t+k,t}\right]\right)^2 .\end{aligned}$$

This result makes clear that MSE depends only on the second moment structure of the joint distribution of the actual and forecasted series. Thus, as noted in Murphy and Winkler (1987, 1992), although MSE is a useful summary statistic for the joint distribution of y_{t+k} and $\hat{y}_{t+k,t}$, in general it contains substantially less information than the actual joint distribution itself. Other statistics highlighting different aspects of the joint distribution may therefore be useful as well. Ultimately, of course, one may want to focus directly on estimates of the joint distribution, which may be available if the sample size is large enough to permit relatively precise estimation.

Measuring forecastability

It is natural and informative to evaluate the accuracy of a forecast. We hasten to add, however, that actual and forecasted values may be dissimilar, even for very good forecasts. To take an extreme example, note that the linear least squares forecast for a zero-mean white noise process is simply zero – the paths of forecasts and realizations will look very different, yet there does not exist a better linear forecast under quadratic loss. This example highlights the inherent limits to forecastability, which depends on the process being forecast; some processes are inherently easy to forecast, while others are hard to forecast. In other words, sometimes the information on which the forecaster optimally conditions is very valuable, and sometimes it isn't.

The issue of how to quantify forecastability arises at once. Granger and Newbold (1976) propose a natural definition of forecastability for covariance stationary series under squared-error loss, patterned after the familiar R^2 of linear regression

$$G = \frac{\mathrm{var}(\hat{y}_{t+1,t})}{\mathrm{var}(y_{t+1})} = 1 - \frac{\mathrm{var}(e_{t+1,t})}{\mathrm{var}(y_{t+1})} ,$$

where both the forecast and forecast error refer to the optimal (that is, linear least squares or conditional mean) forecast.

In closing this section, we note that although measures of forecastability are useful constructs, they are driven by the population properties of processes and their optimal forecasts, so they don't help one to evaluate the "goodness" of an actual reported forecast, which may be far from optimal. For example, if the variance of $\hat{y}_{t+1,t}$ is not much lower than the variance of the covariance stationary series y_{t+1}, it could be that either the forecast is poor, the series is inherently almost unforecastable, or both.

Statistical comparison of forecast accuracy[7]

Once a loss function has been decided upon, it is often of interest to know which of the competing forecasts has smallest expected loss. Forecasts may of course be ranked according to average loss over the sample period, but one would like to have a measure of the sampling variability in such average losses. Alternatively, one would like to be able to test the hypothesis that the difference of expected losses between forecasts *I* and *j* is zero (i.e., $\mathrm{E}[L(y_{t+k}, \hat{y}^i_{t+k,t})] = \mathrm{E}[L(y_{t+k}, \hat{y}^j_{t+k,t})]$), against the alternative that one forecast is better.

[7] This section draws heavily upon Diebold and Mariano (1995).

Stekler (1987) proposes a rank-based test of the hypothesis that each of a set of forecasts has equal expected loss.[8] Given N competing forecasts, assign to each forecast at each time a rank according to its accuracy (the best forecast receives a rank of N, the second-best receives a rank of $N-1$, and so forth). Then aggregate the period-by-period ranks for each forecast,

$$H^i = \sum_{t=1}^{T} \mathrm{Rank}(L(y_{t+k}, \hat{y}^i_{t+k,t})) \ ,$$

$I = 1, \ldots, N$, and form the chi-squared goodness-of-fit test statistic,

$$H = \sum_{i=1}^{N} \frac{(H^i - NT/2)^2}{NT/2} \ .$$

Under the null, $H \sim \chi^2_{N-1}$. As described here, the test requires the rankings to be independent over space and time, but simple modifications along the lines of the Bonferroni bounds test may be made if the rankings are temporally $(k-1)$-dependent. Moreover, exact versions of the test may be obtained by exploiting Fisher's randomization principle.[9]

One limitation of Stekler's rank-based approach is that information on the *magnitude* of differences in expected loss across forecasters is discarded. In many applications, one wants to know not only *whether* the difference of expected losses differs from zero (or the ratio differs from 1), but also by *how much* it differs. Effectively, one wants to know the sampling distribution of the sample mean loss differential (or of the individual sample mean losses), which in addition to being directly informative would enable Wald tests of the hypothesis that the expected loss differential is zero. Diebold and Mariano (1995), building on earlier work by Granger and Newbold (1986) and Meese and Rogoff (1988), develop a test for a zero expected loss differential that allows for forecast errors that are nonzero mean, non-Gaussian, serially correlated and contemporaneously correlated.

In general, the loss function is $L(y_{t+k}, \hat{y}^i_{t+k,t})$. Because in many applications the loss function will be a direct function of the forecast error, $L(y_{t+k}, \hat{y}^i_{t+k,t}) = L(e^i_{t+k,t})$, we write $L(e^i_{t+k,t})$ from this point on to economize on notation, while recognizing that certain loss functions (such as direction-of-change) don't collapse to the $L(e^i_{t+k,t})$ form.[10] The null hypothesis of equal forecast accuracy for two forecasts is $\mathrm{E}[L(e^i_{t+k,t})] = \mathrm{E}[L(e^j_{t+k,t})]$, or $\mathrm{E}[d_t] = 0$, where $d_t \equiv L(e^i_{t+k,t}) - L(e^j_{t+k,t})$ is the loss differential.

If d_t is a covariance stationary, short-memory series, then standard results may be used to deduce the asymptotic distribution of the sample mean loss differential,

$$\sqrt{T}(\bar{d} - \mu) \overset{a}{\sim} N(0, 2\pi f_d(0)) \ ,$$

[8] Stekler uses RMSE, but other loss functions may be used.

[9] See, for example, Bradley (1968), Chapter 4.

[10] In such cases, the $L(Y_{t+k}, \hat{y}_{i,t+k,t})$ form should be used.

where $\bar{d} = 1/T\sum_{t=1}^{T}\left[L(e^i_{t+k,t}) - L(e^j_{t+k,t})\right]$ is the sample mean loss differential, $f_d(0) = 1/2\pi\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$ is the spectral density of the loss differential at frequency zero, $\gamma_d(\tau) = \mathrm{E}[(d_t - \mu)(d_{t-\tau} - \mu)]$ is the autocovariance of the loss differential at displacement τ, and μ is the population mean loss differential. The formula for $f_d(0)$ shows that the correction for serial correlation can be substantial, even if the loss differential is only weakly serially correlated, due to the cumulation of the autocovariance terms. In large samples, the obvious statistic for testing the null hypothesis of equal forecast accuracy is the standardized sample mean loss differential,

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

where $\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$.

It is useful to have available exact finite-sample tests of forecast accuracy to complement the asymptotic tests. As usual, variants of the sign and signed-rank tests are applicable. When using the sign test, the null hypothesis is that the median of the loss differential is zero, $\text{median}(L(e^i_{t+k,t}) - L(e^j_{t+k,t})) = 0$. Note that the null of a zero median loss differential is not the same as the null of zero difference between median losses; that is, $\text{median}(L(e^i_{t+k,t}) - L(e^j_{t+k,t})) \neq \text{median}(L(e^i_{t+k,t})) - \text{median}(L(e^j_{t+k,t}))$. For this reason, the null differs slightly in spirit from that associated with the asymptotic Diebold-Mariano test, but nevertheless, it has the intuitive and meaningful interpretation that $P(L(e^i_{t+k,t}) > L(e^j_{t+k,t})) = P(L(e^i_{t+k,t}) < L(e^j_{t+k,t}))$.

When using the Wilcoxon signed-rank test, the null hypothesis is that the loss differential series is symmetric about a zero median (and hence mean), which corresponds precisely to the null of the asymptotic Diebold-Mariano test. Symmetry of the loss differential will obtain, for example, if the distributions of $L(e^i_{t+k,t})$ and $L(e^j_{t+k,t})$ are the same up to a location shift. Symmetry is ultimately an empirical matter and may be assessed using standard procedures.

The construction and intuition of the distribution-free nonparametric test statistics are straightforward. The sign test statistic is $S_B = \sum_{t=1}^{T} I_+(d_t)$, and the signed-rank test statistic is $W_B = \sum_{t=1}^{T} I_+(d_t)\text{Rank}(|d_t|)$. Serial correlation may be handled as before via Bonferroni bounds. It is interesting to note that, in multistep forecast comparisons, forecast error serial correlation may be a "common feature" in the terminology of Engle and Kozicki (1993), because it is induced largely by the fact that the forecast horizon is longer than the interval at which the data are sampled and may therefore not be present in loss *differentials* even if present in the forecast errors themselves. This possibility can of course be checked empirically.

West (1994) takes an approach very much related to, but nevertheless different from, that of Diebold and Mariano. The main difference is that West assumes that forecasts are computed from an estimated regression model and explicitly accounts for the effects of parameter uncertainty within that framework. When the estimation sample is small, the tests can lead to different results. However, as the estimation period grows in length relative to the forecast period, the effects of parameter uncertainty vanish, and the Diebold-Mariano and West statistics are identical.

West's approach is both more general and less general than the Diebold-Mariano approach. It is more general in that it corrects for nonstationarities induced by the updating of parameter estimates. It is less general in that those corrections are made within the confines of a more rigid framework than that of Diebold and Mariano, in whose framework no assumptions need be made about the often unknown or incompletely known models that underlie forecasts.

In closing this section, we note that it is sometimes informative to compare the accuracy of a forecast to that of a "naive" competitor. A simple and popular such comparison is achieved by Theil's (1961) U statistic, which is the ratio of the 1-step-ahead MSE for a given forecast relative to that of a random walk forecast $\hat{y}_{t+1,t} = y_t$; that is,

$$U = \frac{\sum_{t=1}^{T}\left(y_{t+1} - \hat{y}_{t+1,t}\right)^2}{\sum_{t=1}^{T}(y_{t+1} - y_t)^2} .$$

Generalization to other loss functions and other horizons is immediate. The statistical significance of the MSE comparison underlying the U statistic may be ascertained using the methods just described. One must remember, of course, that the random walk is not necessarily a naive competitor, particularly for many economic and financial variables, so that values of the U statistic near one are not necessarily "bad." Several authors, including Armstrong and Fildes (1995), have advocated using the U statistic and close relatives for comparing the accuracy of various forecasting methods across series.

3. Combining forecasts

In forecast accuracy comparison, one asks which forecast is best with respect to a particular loss function. Regardless of whether one forecast is "best," however, the question arises as to whether competing forecasts may be fruitfully combined – in similar fashion to the construction of an asset portfolio – to produce a composite forecast superior to all the original forecasts. Thus, forecast combination, although obviously related to forecast accuracy comparison, is logically distinct and of independent interest.

Forecast encompassing tests

Forecast encompassing tests enable one to determine whether a certain forecast incorporates (or enczompasses) all the relevant information in competing forecasts. The idea dates at least to Nelson (1972) and Cooper and Nelson (1975), and was formalized and extended by Chong and Hendry (1986). For simplicity, let us focus on the case of two forecasts, $\hat{y}^1_{t+k,t}$ and $\hat{y}^2_{t+k,t}$. Consider the regression

$$y_{t+k} = \beta_0 + \beta_1 \hat{y}^1_{t+k,t} + \beta_2 \hat{y}^2_{t+k,t} + \varepsilon_{t+k,t} \ .$$

If $(\beta_0, \beta_1, \beta_2) = (0, 1, 0)$, one says that model 1 forecast-encompasses model 2, and if $(\beta_0, \beta_1, \beta_2) = (0, 0, 1)$, then model 2 forecast-encompasses model 1. For any other $(\beta_0, \beta_1, \beta_2)$ values, neither model encompasses the other, and both forecasts contain useful information about y_{t+k}. Under certain conditions, the encompassing hypotheses can be tested using standard methods.[11] Moreover, although it does not yet seem to have appeared in the forecasting literature, it would be straightforward to develop exact finite-sample tests (or bounds tests when $k > 1$) of the hypothesis using simple generalizations of the distribution-free tests discussed earlier.

Fair and Shiller (1989, 1990) take a different but related approach based on the regression

$$(y_{t+k} - y_t) = \beta_0 + \beta_1 \left(\hat{y}^1_{t+k,t} - y_t\right) + \beta_2 \left(\hat{y}^2_{t+k,t} - y_t\right) + \varepsilon_{t+k,t} \ .$$

As before, forecast-encompassing corresponds to coefficient values of (0,1,0) or (0,0,1). Under the null of forecast encompassing, the Chong-Hendry and Fair-Shiller regressions are identical. When the variable being forecast is integrated, however, the Fair-Shiller framework may prove more convenient, because the specification in terms of changes facilitates the use of Gaussian asymptotic distribution theory.

Forecast combination

Failure of one model's forecasts to encompass other models' forecasts indicates that all the models examined are misspecified. It should come as no surprise that such situations are typical in practice, because all forecasting models are surely misspecified – they are intentional abstractions of a much more complex reality. What, then, is the role of forecast combination techniques? In a world in which information sets can be instantaneously and costlessly combined, there is no role; it is always optimal to combine information sets rather than forecasts. In the long run, the combination of information sets may sometimes be achieved by improved model specification. But in the short run – particularly when deadlines must be met and timely forecasts produced – pooling of information sets is typically either impossible or prohibitively costly. This simple insight motivates the pragmatic idea of forecast combination, in which forecasts rather than models are the basic object of analysis, due to an assumed inability to combine information sets. Thus, forecast combination can be viewed as a key link between the short-

[11] Note that MA($k-1$) serial correlation will typically be present in $\varepsilon_{t+k,t}$ if $k > 1$.

run, real-time forecast production process, and the longer-run, ongoing process of model development.

Many combining methods have been proposed, and they fall roughly into two groups, "variance-covariance" methods and "regression-based" methods. Let us consider first the variance-covariance method due to Bates and Granger (1969). Suppose one has two unbiased forecasts from which a composite is formed as[12]

$$\hat{y}^c_{t+k,t} = \omega \hat{y}^1_{t+k,t} + (1-\omega)\hat{y}^2_{t+k,t} \ .$$

Because the weights sum to unity, the composite forecast will necessarily be unbiased. Moreover, the combined forecast error will satisfy the same relation as the combined forecast; that is,

$$e^c_{t+k,t} = \omega e^1_{t+k,t} + (1-\omega)e^2_{t+k,t} \ ,$$

with a variance $\sigma^2_c = \omega^2\sigma^2_{11} + (1-\omega)^2\sigma^2_{22} + 2\omega(1-\omega)\sigma_{12}$, where σ^2_{11} and σ^2_{22} are unconditional forecast error variances and σ_{12} is their covariance. The combining weight that minimizes the combined forecast error variance (and hence the combined forecast error MSE, by unbiasedness) is

$$\omega^* = \frac{\sigma^2_{22} - \sigma_{12}}{\sigma^2_{22} + \sigma^2_{11} - 2\sigma_{12}} \ .$$

Note that the optimal weight is determined by both the underlying variances and covariances. Moreover, it is straightforward to show that, except in the case where one forecast encompasses the other, the forecast error variance from the optimal composite is less than $\min(\sigma^2_{11}, \sigma^2_{22})$. Thus, in population, one has nothing to lose by combining forecasts and potentially much to gain.

In practice, one replaces the unknown variances and covariances that underlie the optimal combining weights with consistent estimates; that is, one estimates ω^* by replacing σ_{ij} with $\hat{\sigma}_{ij} = 1/T\sum_{t=1}^{T} e^i_{t+k,t}e^j_{t+k,t}$, yielding

$$\hat{\omega}^* = \frac{\hat{\sigma}^2_{22} - \hat{\sigma}_{12}}{\hat{\sigma}^2_{22} + \hat{\sigma}^2_{11} - 2\hat{\sigma}_{12}} \ .$$

In finite samples of the size typically available, sampling error contaminates the combining weight estimates, and the problem of sampling error is exacerbated by the collinearity that typically exists among primary forecasts. Thus, while one hopes to reduce out-of-sample forecast MSE by combining, there is no guarantee. In practice, however, it turns out that forecast combination techniques often perform very well, as documented Clemen's (1989) review of the vast literature on forecast combination.

Now consider the "regression method" of forecast combination. The form of the Chong-Hendry and Fair-Shiller encompassing regressions immediately sug-

[12] The generalization to the case of $M > 2$ competing unbiased forecasts is straightforward, as shown in Newbold and Granger (1974).

gests combining forecasts by simply regressing realizations on forecasts. Granger and Ramanathan (1984) showed that the optimal variance-covariance combining weight vector has a regression interpretation as the coefficient vector of a linear projection of y_{t+k} onto the forecasts, subject to two constraints: the weights sum to unity, and no intercept is included. In practice, of course, one simply runs the regression on available data.

In general, the regression method is simple and flexible. There are many variations and extensions, because any "regression tool" is potentially applicable. The key is to use generalizations with sound motivation. We shall give four examples: time-varying combining weights, dynamic combining regressions, Bayesian shrinkage of combining weights toward equality, and nonlinear combining regressions.

a. Time-varying combining weights

Time-varying combining weights were proposed in the variance-covariance context by Granger and Newbold (1973) and in the regression context by Diebold and Pauly (1987). In the regression framework, for example, one may undertake weighted or rolling estimation of combining regressions, or one may estimate combining regressions with explicitly time-varying parameters.

The potential desirability of time-varying weights stems from a number of sources. First, different learning speeds may lead to a particular forecast improving over time relative to others. In such situations, one naturally wants to weight the improving forecast progressively more heavily. Second, the design of various forecasting models may make them relatively better forecasting tools in some situations than in others. For example, a structural model with a highly developed wage-price sector may substantially outperform a simpler model during times of high inflation. In such times, the more sophisticated model should received higher weight. Third, the parameters in agents' decision rules may drift over time, and certain forecasting techniques may be relatively more vulnerable to such drift.

b. Dynamic combining regressions

Serially correlated errors arise naturally in combining regressions. Diebold (1988) considers the covariance stationary case and argues that serial correlation is likely to appear in unrestricted regression-based forecast combining regressions when $\beta_1 + \beta_2 \neq 1$. More generally, it may be a good idea to allow for serial correlation in combining regressions to capture any dynamics in the variable to be forecast not captured by the various forecasts. In that regard, Coulson and Robins (1993), following Hendry and Mizon (1978), point out that a combining regression with serially correlated disturbances is a special case of a combining regression that includes lagged dependent variables and lagged forecasts, which they advocate.

c. Bayesian shrinkage of combining weights toward equality

Simple arithmetic averages of forecasts are often found to perform very well, even relative to "optimal" composites.[13] Obviously, the imposition of an equal weights constraint eliminates variation in the estimated weights at the cost of possibly introducing bias. However, the evidence indicates that, under quadratic loss, the benefits of imposing equal weights often exceed this cost. With this in mind, Clemen and Winkler (1986) and Diebold and Pauly (1990) propose Bayesian shrinkage techniques to allow for the incorporation of varying degrees of prior information in the estimation of combining weights; least-squares weights and the prior weights then emerge as polar cases for the posterior-mean combining weights. The actual posterior mean combining weights are a matrix weighted average of those for the two polar cases. For example, using a natural conjugate normal-gamma prior, the posterior-mean combining weight vector is

$$\beta^{\text{posterior}} = (Q + F'F)^{-1}(Q\beta^{\text{prior}} + F'F\beta) \ ,$$

where β^{prior} is the prior mean vector, Q is the prior precision matrix, F is the design matrix for the combining regression, and $\hat{\beta}$ is the vector of least squares combining weights. The obvious shrinkage direction is toward a measure of central tendency (e.g., the arithmetic mean). In this way, the combining weights are coaxed toward the arithmetic mean, but the data are still allowed to speak, when (and if) they have something to say.

d. Nonlinear combining regressions

There is no reason, of course, to force combining regressions to be linear, and various of the usual alternatives may be entertained. One particularly interesting possibility is proposed by Deutsch, Granger and Teräsvirta (1994), who suggest

$$\begin{aligned}\hat{y}^c_{t+k,t} &= I(s_t = 1)\left(\beta_{11}\hat{y}^1_{t+k,t} + \beta_{12}\hat{y}^2_{t+k,t}\right) \\ &\quad + I(s_t = 2)\left(\beta_{21}\hat{y}^1_{t+k,t} + \beta_{22}\hat{y}^2_{t+k,t}\right) \ .\end{aligned}$$

The states that govern the combining weights can depend on past forecast errors from one or both models or on various economic variables. Furthermore, the indicator weight need not be simply a binary variable; the transition between states can be made more gradual by allowing weights to be functions of the forecast errors or economic variables.

4. Special topics in evaluating economic and financial forecasts

Evaluating direction-of-change forecasts

Direction-of-change forecasts are often used in financial and economic decision-making (e.g., Leitch and Tanner, 1991, 1995; Satchell and Timmermann, 1992).

[13] See Winkler and Makridakis (1983), Cleman (1989), and many of the references therein.

The question of how to evaluate such forecasts immediately arises. Our earlier results on tests for forecast accuracy comparison remain valid, appropriately modified, so we shall not restate them here. Instead, we note that one frequently sees assessments of whether direction-of-change forecasts "have value," and we shall discuss that issue.

The question as to whether a direction-of-change forecast has value by necessity involves comparison to a naive benchmark – the direction-of-change forecast is compared to a "naive" coin flip (with success probability equal to the relevant marginal). Consider a 2 × 2 contingency table. For ease of notation, call the two states into which forecasts and realizations fall "*I*" and "*j*". Commonly, for example, I = "up" and j = "down." Tables 1 and 2 make clear our notation regarding observed cell counts and unobserved cell probabilities. The null hypothesis that a direction-of-change forecast has no value is that the forecasts and realizations are independent, in which case $P_{ij} = P_{i.}P_{.j}$, $\forall i,\ j$. As always, one proceeds under the null. The true cell probabilities are of course unknown, so one uses the consistent estimates $\hat{P}_{i.} = O_{i.}/O$ and $\hat{P}_{.j} = O_{.j}/O$. Then one consistently estimates the expected cell counts under the null, $\mathrm{E}_{ij} = P_{i.}P_{.j}O$, by $\hat{E}_{ij} = \hat{P}_{i.}\hat{P}_{.j}O = O_{i.}O_{.j}/O$. Finally, one constructs the statistic $C = \sum_{i,j=1}^{2}(O_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$. Under the null, $C \xrightarrow{d} \chi^2_1$.

An intimately-related test of forecast value was proposed by Merton (1981) and Henriksson and Merton (1981), who assert that a forecast has value if $P_{ii}/P_{i.} + P_{jj}/P_{j.} > 1$. They therefore develop an exact test of the null hypothesis that $P_{ii}/P_{i.} + P_{jj}/P_{j.} = 1$ against the inequality alternative. A key insight, noted in varying degrees by Schnader and Stekler (1990) and Stekler (1994), and formalized by Pesaran and Timmermann (1992), is that the Henriksson-Merton null is equivalent to the contingency-table null if the marginal probabilities are fixed at the observed relative frequencies, $O_{i.}/O$ and $O_{.j}/O$. The same unpalatable assumption is necessary for deriving the exact finite-sample distribution of the Henriksson-Merton test statistic.

Table 1
Observed cell counts

	Actual *i*	Actual *j*	Marginal
Forecast *i*	O_{ii}	O_{ij}	$O_{i.}$
Forecast *j*	O_{ji}	O_{jj}	O_{j}
Marginal	$O_{.i}$	$O_{.j}$	Total: O

Table 2
Unobserved cell probabilities

	Actual *i*	Actual *j*	Marginal
Forecast *i*	P_{ii}	P_{ij}	$P_{i.}$
Forecast *j*	P_{ji}	P_{jj}	$P_{j.}$
Marginal	$P_{.i}$	$P_{.j}$	Total: 1

Asymptotically, however, all is well; the square of the Henriksson-Merton statistic, appropriately normalized, is asymptotically equivalent to C, the chi-squared contingency table statistic. Moreover, the 2 × 2 contingency table test generalizes trivially to the $N \times N$ case, with

$$C_N = \sum_{i,j=1}^{N} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \quad .$$

Under the null, $C_N \overset{a}{\sim} \chi^2_{(N-1)(N-1)}$. A subtle point arises, however, as pointed out by Pesaran and Timmermann (1992). In the 2 × 2 case, one must base the test on the entire table, as the off-diagonal elements are determined by the diagonal elements, because the two elements of each row must sum to one. In the $N \times N$ case, in contrast, there is more latitude as to which cells to examine, and for purposes of forecast evaluation, it may be desirable to focus only on the diagonal cells.

In closing this section, we note that although the contingency table tests are often of interest in the direction-of-change context (for the same reason that tests based on Theil's U-statistic are often of interest in more standard contexts), forecast "value" in that sense is neither a necessary nor sufficient condition for forecast value in terms of a profitable trading strategy yielding significant excess returns. For example, one might beat the marginal forecast but still earn no excess returns after adjusting for transactions costs. Alternatively, one might do worse than the marginal but still make huge profits if the "hits" are "big," a point stressed by Cumby and Modest (1987).

Evaluating probability forecasts

Oftentimes economic and financial forecasts are issued as probabilities, such as the probability that a business cycle turning point will occur in the next year, the probability that a corporation will default on a particular bond issue this year, or the probability that the return on the S&P 500 stock index will be more than ten percent this year. A number of specialized considerations arise in the evaluation of probability forecasts, to which we now turn. Let $P_{t+k,t}$ be a probability forecast made at time t for an event at time $t+k$, and let $R_{t+k} = 1$ if the event occurs and zero otherwise. $P_{t+k,t}$ is a scalar if there are only two possible events. More generally, if there are N possible events, then $P_{t+k,t}$ is an $(N-1) \times 1$ vector.[14] For notational economy, we shall focus on scalar probability forecasts.

Accuracy measures for probability forecasts are commonly called "scores," and the most common is Brier's (1950) quadratic probability score, also called the Brier score,

[14] The probabillity forecast assigned to the N^th^ event is implicitly determined by the restriction that the probabilities sum to 1.

$$\mathrm{QPS} = \frac{1}{T}\sum_{t=1}^{T} 2\left(P_{t+k,t} - R_{t+k}\right)^2 \ .$$

Clearly, QPS $\in [0,2]$, and it has a negative orientation (smaller values indicate more accurate forecasts).[15] To understand the QPS, note that the accuracy of *any* forecast refers to the expected loss when using that forecast, and typically loss depends on the deviation between forecasts and realizations. It seems reasonable, then, in the context of probability forecasting under quadratic loss, to track the average squared divergence between $P_{t+k,t}$ and R_{t+k}, which is what the QPS does. Thus, the QPS is a rough probability-forecast analog of MSE.

The QPS is only a *rough* analog of MSE, however, because $P_{t+k,t}$ is in fact not a *forecast* of the *outcome* (which is 0–1), but rather a probability assigned to it. A more natural and direct way to evaluate probability forecasts is simply to compare the forecasted probabilities to observed relative frequencies – that is, to assess calibration. An overall measure of calibration is the global squared bias,

$$\mathrm{GSB} = 2(\bar{P} - \bar{R})^2 \ ,$$

where $\bar{P} = 1/T\sum_{t=1}^{T} P_{t+k,t}$ and $\bar{R} = 1/T\sum_{t=1}^{T} R_{t+k}$. GSB $\in [0,2]$ with a negative orientation.

Calibration may also be examined locally in any subset of the unit interval. For example, one might check whether the observed relative frequency corresponding to probability forecasts between 0.6 and 0.7 is also between 0.6 and 0.7. One may go farther to form a weighted average of local calibration across all cells of a J-subset partition of the unit interval into J subsets chosen according to the user's interest and the specifics of the situation.[16] This leads to the local squared bias measure,

$$\mathrm{LSB} = \frac{1}{T}\sum_{j=1}^{J} 2T_j\left(\bar{P}_j - \bar{R}_j\right)^2 \ ,$$

where T_j is the number of probability forecasts in set $j, \bar{P}_j$ is the average forecast in set j, and $\bar{R}_j$ is the average realization in set j, j = 1, ..., J. Note that LSB $\in [0,2]$, and LSB = 0 implies that GSB = 0, but not conversely.

Testing for adequate calibration is a straightforward matter, at least under independence of the realizations. For a given event and a corresponding sequence of forecasted probabilities $\{P_{t+k,t}\}_{t=1}^{T}$, create J mutually exclusive and collectively exhaustive subsets of forecasts, and denote the midpoint of each range $\pi_j, j = 1, \ldots, J$. Let R_j denote the number of observed events when the forecast was in set j, respectively, and define "range j" calibration statistics,

[15] The "2" that appears in the QPS formula is an artifact from the full vector case. We could of course drop it without affecting the QPS rankings of competing forecasts, but we leave it to maintain comparability to other literature.

[16] For example, Diebold and Rudebusch (1989) split the unit interval into ten equal parts.

$$Z_j = \frac{(R_j - T_j\pi_j)}{(T_j\pi_j(1-\pi_j))^{1/2}} \equiv \frac{(R_j - e_j)}{w_j^{1/2}}, \; j = 1,\ldots,J \;,$$

and an overall calibration statistic,

$$Z_0 = \frac{(R_+ - e_+)}{w_+^{1/2}} \;,$$

where $R_+ = \sum_{j=1}^J R_j$, $e_+ = \sum_{j=1}^J T_j\pi_j$, and $w_+ = \sum_{j=1}^J T_j\pi_j(1-\pi_j)$. Z_0 is a joint test of adequate local calibration across all cells, while the Z_j statistics test cell-by-cell local calibration.[17] Under independence, the binomial structure would obviously imply that $Z_0 \overset{a}{\sim} N(0,1)$, and $Z_j \overset{a}{\sim} N(0,1), \forall j = 1,\ldots,J$. In a fascinating development, Seillier-Moiseiwitsch and Dawid (1993) show that the asymptotic normality holds much more generally, including in the dependent situations of practical relevance.

One additional feature of probability forecasts (or more precisely, of the corresponding realizations), called resolution, is of interest:

$$\text{RES} = \frac{1}{T}\sum_{j=1}^{J} 2T_j(\bar{R}_j - \bar{R})^2 \;.$$

RES is simply the weighted average squared divergence between $\bar{R}$ and the $\bar{R}'_j s$, a measure of how much the observed relative frequencies move across cells. RES ≥ 0 and has a positive orientation. As shown by Murphy (1973), an informative decomposition of QPS exists,

$$\text{QPS} = \text{QPS}_{\bar{R}} + \text{LSB} - \text{RES} \;,$$

where $\text{QPS}_{\bar{R}}$ is the QPS evaluated at $P_{t+k,t} = \bar{R}$. This decomposition highlights the tradeoffs between the various attributes of probability forecasts.

Just as with Theil's U-statistic for "standard" forecasts, it is sometimes informative to compare the performance of a particular probability forecast to that of a benchmark. Murphy (1974), for example, proposes the statistic

$$M = \text{QPS} - \text{QPS}_{\bar{R}} = \text{LSB} - \text{RES} \;,$$

which measures the difference in accuracy between the forecast at hand and the benchmark forecast $\bar{R}$. Using the earlier-discussed Diebold-Mariano approach, one can also assess the *significance* of differences in QPS and $\text{QPS}_{\bar{R}}$, differences in QPS or various other measures of probability forecast accuracy across forecasters, or differences in local or global calibration across forecasters.

[17] One may of course test for adequate global calibration by using a trivial partition of the unit interval – the unit interval itself.

Evaluating volatility forecasts

Many interesting questions in finance, such as options pricing, risk hedging and portfolio management, explicitly depend upon the variances of asset prices. Thus, a variety of methods have been proposed for generating volatility forecasts. As opposed to point or probability forecasts, evaluation of volatility forecasts is complicated by the fact that actual conditional variances are *unobservable*.

A standard "solution" to this unobservability problem is to use the squared realization ε_{t+k}^2 as a proxy for the true conditional variance h_{t+k}, because $\mathrm{E}[\varepsilon_{t+k}^2|\Omega_{t+k-1}] = \mathrm{E}[h_{t+k}v_{t+k}^2|\Omega_{t+k-1}] = h_{t+k}$, where $v_{t+k} \sim WN(0,1)$.[18] Thus, for example, $\mathrm{MSE} = 1/T\sum_{t=1}^{T}(\varepsilon_{t+k}^2 - \hat{h}_{t+k,t})^2$. Although MSE is often used to measure volatility forecast accuracy, Bollerslev, Engle and Nelson (1994) point out that MSE is inappropriate, because it penalizes positive volatility forecasts and negative volatility forecasts (which are meaningless) symmetrically. Two alternative loss functions that penalize volatility forecasts asymmetrically are the logarithmic loss function employed in Pagan and Schwert (1990),

$$\mathrm{LL} = \frac{1}{T}\sum_{t=1}^{T}\left[\ln(\varepsilon_{t+k}^2) - \ln\left(\hat{h}_{t+k,t}\right)\right]^2 ,$$

and the heteroskedasticity-adjusted MSE of Bollerslev and Ghysels (1994),

$$\mathrm{HMSE} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\varepsilon_{t+k}^2}{\hat{h}_{t+k,t}} - 1\right]^2 .$$

Bollerslev, Engle and Nelson (1994) suggest the loss function implicit in the Gaussian quasi-maximum likelihood function often used in fitting volatility models; that is,

$$\mathrm{GMLE} = \frac{1}{T}\sum_{t=1}^{T}\left[\ln\left(\hat{h}_{t+k,t}\right) + \frac{\varepsilon_{t+k}^2}{\hat{h}_{t+k,t}}\right] .$$

As with all forecast evaluations, the volatility forecast evaluations of most interest to forecast users are those conducted under the relevant loss function. West, Edison and Cho (1993) and Engle et al. (1993) make important contributions along those lines, proposing economic loss functions based on utility maximization and profit maximization, respectively. Lopez (1995) proposes a framework for volatility forecast evaluation that allows for a variety of economic loss functions. The framework is based on transforming volatility forecasts into probability forecasts by integrating over the assumed or estimated distribution of ε_t. By selecting the range of integration corresponding to an event of interest, a

[18] Although ε_{t+k}^2 is an unbiased estimator of h_{t+k}, it is an imprecise or "noisy" estimator. For example, if $v_{t+k} \sim N(0,1)$,$\varepsilon_{t+k}^2 = h_{t+k}v_{t+k}^2$ has a conditional mean of h_{t+k} because $v_{t+k}^2 \sim \chi_1^2$. Yet, because the median of a χ_1^2 distribution is 0.455, $\varepsilon_{t+k}^2 < 1/2h_{t+k}$ more than fifty percent of the time.

forecast user can incorporate elements of her loss function into the probability forecasts.

For example, given $\varepsilon_{t+k}|\Omega_t \sim D(0, h_{t+k,t})$ and a volatility forecast $\hat{h}_{t+k,t}$, an options trader interested in the event $\varepsilon_{t+k} \in [L_{\varepsilon,t+k}, U_{\varepsilon,t+k}]$ would generate the probability forecast

$$\begin{aligned} P_{t+k,t} &= Pr\left(L_{\varepsilon,t+k} < \varepsilon_{t+k} < U_{\varepsilon,t+k}\right) \\ &= Pr\left(\frac{L_{\varepsilon,t+k}}{\sqrt{\hat{h}_{t+k,t}}} < z_{t+k} < \frac{U_{\varepsilon,t+k}}{\sqrt{\hat{h}_{t+k,t}}}\right) = \int_{l_{\varepsilon,t+k}}^{u_{\varepsilon,t+k}} f(z_{t+k}) dz_{t+k} \ , \end{aligned}$$

where z_{t+k} is the standardized innovation, $f(z_{t+k})$ is the functional form of $D(0, 1)$, and $\left[l_{\varepsilon,t+k}, u_{\varepsilon,t+k}\right]$ is the standardized range of integration. In contrast, a forecast user interested in the behavior of the underlying asset, $y_{t+k} = \mu_{t+k,t} + \varepsilon_{t+k}$ where $\mu_{t+k,t} = \mathrm{E}[y_{t+k}|\Omega_t]$, might generate the probability forecast

$$\begin{aligned} P_{t+k,t} &= Pr\left(L_{y,t+k} < y_{t+k} < U_{y,t+k}\right) \\ &= Pr\left(\frac{L_{y,t+k} - \hat{\mu}_{t+k,t}}{\sqrt{\hat{h}_{t+k,t}}} < z_{t+k} < \frac{U_{y,t+k} - \hat{\mu}_{t+k,t}}{\sqrt{\hat{h}_{t+k,t}}}\right) \\ &= \int_{l_{y,t+k}}^{u_{y,t+k}} f(z_{t+k}) dz_{t+k} \ , \end{aligned}$$

where $\hat{\mu}_{t+k,t}$ is the forecasted conditional mean and $\left[l_{y,t+k}, u_{y,t+k}\right]$ is the standardized range of integration.

Once generated, these probability forecasts can be evaluated using the scoring rules described above, and the significance of differences across models can be tested using the Diebold-Mariano tests. The key advantage of this framework is that it allows the evaluation to be based on observable events and thus avoids proxying for the unobservable true variance.

The Lopez approach to volatility forecast evaluation is based on time-varying probabilities assigned to a fixed interval. Alternatively, one may fix the probabilities and vary the widths of the intervals, as in traditional confidence interval construction. In that regard, Christoffersen (1995) suggests exploiting the fact that if a $(1-\alpha)\%$ confidence interval (denoted $[L_{y,t+k}, U_{y,t+k}]$) is correctly calibrated, then

$$\mathrm{E}[I_{t+k,t}|I_{t,t-k}, I_{t-1,t-k-1}, \ldots I_{k+1,1}] = (1-\alpha) \ ,$$

where

$$I_{t+k,t} = \begin{cases} 1, & \text{if } y_{t+k} \in [L_{y,t+k}, U_{y,t+k}] \\ 0, & \text{if otherwise.} \end{cases}$$

That is, Christoffersen suggests checking conditional coverage.[19]

Standard evaluation methods for interval forecasts typically restrict attention to unconditional coverage, $\mathrm{E}[I_{t+k|t}] = (1-\alpha)$. But simply checking unconditional coverage is insufficient in general, because an interval forecast with correct unconditional coverage may nevertheless have incorrect *conditional* coverage at any particular time.

For one-step-ahead interval forecasts ($k=1$), the conditional coverage criterion becomes

$$\mathrm{E}[I_{t+1,t}|I_{t,t-1}, I_{t-1,t-2}, \ldots I_{2,1}] = (1-\alpha) \ ,$$

or equivalently,

$$I_{t+1|t} \overset{iid}{\sim} \mathrm{Bern}(1-\alpha) \ .$$

Given T values of the indicator variable for T interval forecasts, one can determine whether the forecast intervals display correct conditional coverage by testing the hypothesis that the indicator variable is an iid Bernoulli$(1-\alpha)$ random variable. A likelihood ratio test of the iid Bernoulli hypothesis is readily constructed by comparing the log likelihoods of restricted and unrestricted Markov processes for the indicator series $\{I_{t+1,t}\}$. The unrestricted transition probability matrix is

$$\Pi = \begin{bmatrix} \pi_{11} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{00} \end{bmatrix},$$

where $\pi_{11} = P(I_{t+1|t} = 1 | I_{t|t-1} = 1)$, and so forth. The transition probability matrix under the null is $\left[\begin{smallmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{smallmatrix}\right]$ The corresponding approximate likelihood functions are

$$L(\Pi|I) = (\pi_{11})^{n_{11}}(1-\pi_{11})^{n_{10}}(1-\pi_{00})^{n_{01}}\pi_{00}^{n_{00}}$$

and

$$L(\alpha|I) = (1-\alpha)^{(n_{11}+n_{01})}(\alpha)^{(n_{10}+n_{00})} \ ,$$

where n_{ij} is the number of observed transitions from I to j and I is the indicator sequence.[20] The likelihood ratio statistic for the conditional coverage hypothesis is

$$LR_{cc} = 2[lnL(\hat{\Pi}|I) - lnL(\alpha|I)] \ ,$$

[19] In general, one wants to test whether $\mathrm{E}[I_{t+k|t}|\Omega_t] = (1-\alpha)$, where Ω_t is all information available at time t. For present purposes, Ω_t is restricted to past values of the indicator sequence in order to construct general and easily applied tests.

[20] The likelihoods are approximate because the initial terms are dropped. All the likelihood ratio tests presented are of course asymptotic, so the treatment of the initial terms is inconsequential.

where $\hat{\Pi}$ are the maximum likelihood estimates. Under the null hypothesis, $LR_{cc} \overset{a}{\sim} \chi_2^2$.

The likelihood ratio test of conditional coverage can be decomposed into two separately interesting hypotheses, correct unconditional coverage, $\mathrm{E}[I_{t+1|t}] = (1-\alpha)$, and independence, $\pi_{11} = 1 - \pi_{00}$. The likelihood ratio test for correct unconditional coverage (given independence) is

$$LR_{uc} = 2[\ln L(\hat{\pi}|I) - \ln L(\alpha|I)] \ ,$$

where $L(\pi|I) = (1-\pi)^{(n_{11}+n_{01})}(\pi)^{(n_{10}+n_{00})}$. Under the null hypothesis, $LR_{uc} \overset{a}{\sim} \chi_1^2$. The independence hypothesis is tested separately by

$$LR_{\mathrm{ind}} = 2[\ln L(\hat{\Pi}|I) - \ln L(\hat{\pi}|I)] \ .$$

Under the null hypothesis, $LR_{\mathrm{ind}} \overset{a}{\sim} \chi_1^2$. It is apparent that $LR_{cc} = LR_{uc} + LR_{\mathrm{ind}}$, in small as well as large samples.

The independence property can also be checked in the case where $k = 1$ using the group test of David (1947), which is an exact and uniformly most powerful test against first-order dependence. Define a group as a string of consecutive zeros or ones, and let r be the number of groups in the sequence $\{I_{t+1,t}\}$. Under the null that the sequence is iid, the distribution of r given the total number of ones, n_1, and the total number of zeros, n_0, is

$$P(r|n_0, n_1) = \frac{f_r}{\binom{n}{n_0}}, \text{ for } r \geq 2 \ ,$$

where $n = n_0 + n_1$, and

$$f_r = \begin{cases} f_{2s} & = 2\binom{n_0 - 1}{s-1}\binom{n_1 - 1}{s-1}, & \text{for } r \text{ even} \\ f_{2s+1} & = \dfrac{f_{2s}(n-2s)}{(2s)}, & \text{for } r \text{ odd} \ . \end{cases}$$

Finally, the generalization to $k > 1$ is simple in the likelihood ratio framework, in spite of the fact that k-step-ahead prediction errors are serially correlated in general. The basic framework remains intact but requires a k^{th}-order Markov chain. A k^{th}-order chain, however, can always be written as a first-order chain with an expanded state space, so that direct analogs of the results for the first-order case apply.

5. Concluding remarks

Three modern themes permeate this survey, so it is worth highlighting them explicitly. The first theme is that various types of forecasts, such as probability forecasts and volatility forecasts, are becoming more integrated into economic and financial decision making, leading to a derived demand for new types of forecast evaluation procedures.

The second theme is the use of exact finite-sample hypothesis tests, typically based on distribution-free nonparametrics. We explicitly sketched such tests in the context of forecast-error unbiasedness, k-dependence, orthogonality to available information, and when more than one forecast is available, in the context of testing equality of expected loss, testing whether a direction-of-change forecast has value, etc.

The third theme is use of the relevant loss function. This idea arose in many places, such as in forecastability measures and forecast accuracy comparison tests, and may readily be introduced in others, such as orthogonality tests, encompassing tests and combining regressions. In fact, an integrated tool kit for estimation, forecasting, and forecast evaluation (and hence model selection and nonnested hypothesis testing) under the relevant loss function is rapidly becoming available; see Weiss and Andersen (1984), Weiss (1995), Diebold and Mariano (1995), Christoffersen and Diebold (1994, 1995), and Diebold, Ohanian and Berkowitz (1995).

References

Armstrong, J. S. and R. Fildes (1995). On the selection of error measures for comparisons among forecasting methods. *J. Forecasting* **14**, 67–71.

Auerbach, A. (1994). The U.S. fiscal problem: Where we are, how we got here and where we're going. NBER Macroeconomics Annual, MIT Press, Cambridge, MA.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Oper. Res. Quart.* **20**, 451–468.

Bollerslev, T., R. F. Engle and D. B. Nelson (1994). ARCH models. In: R. F. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. 4, North-Holland, Amsterdam.

Bollerslev, T. and E. Ghysels (1994). Periodic autoregressive conditional heteroskedasticity. Working Paper No. 178, Department of Finance, Kellogg School of Management, Northwestern University.

Bonham, C. and R. Cohen (1995). Testing the rationality of price forecasts: Comment. *Amer. Econom. Rev.* **85**, 284–289.

Bradley, J. V. (1968). Distribution-free statistical tests. Prentice Hall, Englewood Cliffs, NJ.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **75**, 1–3.

Brown, B. W. and S. Maital (1981). What do economists know? An empirical study of experts' expectations. *Econometrica* **49**, 491–504.

Campbell, B. and J.-M. Dufour (1991 Over-rejections in rational expectations models: A nonparametric approach to the Mankiw-Shapiro problem. *Econom. Lett.* **35**, 285–290.

Campbell, B. and J.-M. Dufour (1995). Exact nonparametric orthogonality and random walk tests. *Rev. Econom. Statist.* **77**, 1–16.

Campbell, B. and E. Ghysels (1995). Federal budget projections: A nonparametric assessment of bias and efficiency. *Rev. Econom. Statist.* **77**, 17–31.

Campbell, J. Y. and N. G. Mankiw (1987). Are output fluctuations transitory? *Quart. J. Econom.* **102**, 857–880.

Chong, Y. Y. and D. F. Hendry (1986). Econometric evaluation of linear macroeconomic models. *Rev. Econom. Stud.* **53**, 671–690.

Christoffersen, P. F. (1995). Predicting uncertainty in the foreign exchange markets. Manuscript, Department of Economics, University of Pennsylvania.

Christoffersen, P. F. and F. X. Diebold (1994). Optimal prediction under asymmetric loss. Technical Working Paper No. 167, National Bureau of Economic Research, Cambridge, MA.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5**, 559–581.

Clemen, R. T. and R. L. Winkler (1986). Combining economic forecasts. *J. Econom. Business Statist.* **4**, 39–46.

Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors. *J. Forecasting* **12**, 617–638.

Cochrane, J. H. (1988). How big is the random walk in GNP? *J. Politic. Econom.* **96**, 893–920.

Cooper, D. M. and C. R. Nelson (1975). The ex-ante prediction performance of the St. Louis and F.R.B.-M.I.T.-Penn econometric models and some results on composite predictors. *J. Money, Credit and Banking* **7**, 1–32.

Coulson, N. E. and R. P. Robins (1993). Forecast combination in a dynamic setting. *J. Forecasting* **12**, 63–67.

Cumby, R. E. and J. Huizinga (1992). Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. *Econometrica* **60**, 185–195.

Cumby, R. E. and D. M. Modest (1987). Testing for market timing ability: A framework for forecast evaluation. *J. Financ. Econom.* **19**, 169–189.

David, F. N. (1947). A power function for tests of randomness in a sequence of alternatives. *Biometrika* **34**, 335–339.

Deutsch, M., C. W. J. Granger and T. Tersvirta (1994). The combination of forecasts using changing weights. *Internat. J. Forecasting* **10**, 47–57.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. *J. Business Econom. Statist.* **6**, 105–111.

Diebold, F. X. (1993). On the limitations of comparing mean square forecast errors: Comment. *J. Forecasting* **12**, 641–642.

Diebold, F. X. and P. Lindner (1995). Fractional integration and interval prediction. *Econom. Lett.*, to appear.

Diebold, F. X. and R. Mariano (1995). Comparing predictive accuracy. *J. Business Econom. Statist.* **13**, 253–264.

Diebold, F. X. L. Ohanian and J. Berkowitz (1995). Dynamic equilibrium economies: A framework for comparing models and data. Technical Working Paper No. 174, National Bureau of Economic Research, Cambridge, MA.

Diebold, F. X. and P. Pauly (1987). Structural change and the combination of forecasts. *J. Forecasting* **6**, 21–40.

Diebold, F. X. and P. Pauly (1990). The use of prior information in forecast combination. *Internat. J. Forecasting* **6**, 503–508.

Diebold, F. X. and G. D. Rudebusch (1989). Scoring the leading indicators. *J. Business* **62**, 369–391.

Dufour, J.-M. (1981). Rank tests for serial dependence. *J. Time Ser. Anal.* **2**, 117–128.

Engle, R. F., C.-H. Hong A. Kane and J. Noh (1993). Arbitrage valuation of variance forecasts with simulated options. In: D. Chance and R. Tripp, eds., Advances in Futures and Options Research, JIA Press, Greenwich, CT.

Engle, R. F. and S. Kozicki (1993). Testing for common features. *J. Business Econom. Statist.* **11**, 369–395.

Fair, R. C. and R. J. Shiller (1989). The informational content of ex-ante forecasts. *Rev. Econom. Statist.* **71**, 325–331.

Fair, R. C. and R. J. Shiller (1990). Comparing information in forecasts from econometric models. *Amer. Econom. Rev.* **80**, 375–389.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *J. Finance* **25**, 383–417.

Fama, E. F. (1975). Short-term interest rates as predictors of inflation. *Amer. Econom. Rev.* **65**, 269–282.

Fama, E. F. (1991). Efficient markets II. *J. Finance* **46**, 1575–1617.

Fama, E. F. and K. R. French (1988). Permanent and temporary components of stock prices. *J. Politic. Econom.* **96**, 246–273.

Granger, C. W. J. and P. Newbold (1973). Some comments on the evaluation of economic forecasts. *Appl. Econom.* **5**, 35–47.

Granger, C. W. J. and P. Newbold (1976). Forecasting transformed series. *J. Roy. Statist. Soc. B* **38**, 189–203.

Granger, C. W. J. and P. Newbold (1986). Forecasting economic time series. 2nd ed., Academic Press, San Diego.

Granger, C. W. J. and R. Ramanathan (1984). Improved methods of forecasting. *J. Forecasting* **3**, 197–204.

Hansen, L. P. and R. J. Hodrick (1980). Forward exchange rates as optimal predictors of future spot rates: An econometric investigation. *J. Politic. Econom.* **88**, 829–853.

Hendry, D. F. and G. E. Mizon (1978). Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. *Econom. J.* **88**, 549–563.

Henriksson, R. D. and R. C. Merton (1981). On market timing and investment performance II: Statistical procedures for evaluating forecast skills. *J. Business* **54**, 513–533.

Keane, M. P. and D. E. Runkle (1990). Testing the rationality of price forecasts: New evidence from panel data. *Amer. Econom. Rev.* **80**, 714–735.

Leitch, G. and J. E. Tanner (1991). Economic forecast evaluation: Profits versus the conventional error measures. *Amer. Econom. Rev.* **81**, 580–590.

Leitch, G. and J. E. Tanner (1995). Professional economic forecasts: Are they worth their costs? *J. Forecasting* **14**, 143–157.

LeRoy, S. F. and R. D. Porter (1981). The present value relation: Tests based on implied variance bounds. *Econometrica* **49**, 555–574.

Lopez, J. A. (1995). Evaluating the predictive accuracy of volatility models. Manuscript, Research and Market Analysis Group, Federal Reserve Bank of New York.

Mark, N. C. (1995). Exchange rates and fundamentals: Evidence on long-horizon predictability. *Amer. Econ. Rev.* **85**, 201–218.

McCulloch, R. and P. E. Rossi (1990). Posterior, predictive and utility-based approaches to testing the arbitrage pricing theory. *J. Financ. Econ.* **28**, 7–38.

Meese, R. A. and K. Rogoff (1988). Was it real? The exchange rate – interest differential relation over the modern floating-rate period. *J. Finance* **43**, 933–948.

Merton, R. C. (1981). On market timing and investment performance I: An equilibrium theory of value for market forecasts. *J. Business* **54**, 513–533.

Mincer, J. and V. Zarnowitz (1969). The evaluation of economic forecasts. In: J. Mincer, ed., Economic forecasts and expectations, National Bureau of Economic Research, New York.

Murphy, A. H. (1973). A new vector partition of the probability score. *J. Appl. Meteor.* **12**, 595–600.

Murphy, A. H. (1974). A sample skill score for probability forecasts. *Monthly Weather Review* **102**, 48–55.

Murphy, A. H. and R. L. Winkler (1987). A general framework for forecast evaluation. *Monthly Weather Review* **115**, 1330–1338.

Murphy, A. H. and R. L. Winkler (1992). Diagnostic verification of probability forecasts. *Internat. J. Forecasting* **7**, 435–455.

Nelson, C. R. (1972). The prediction performance of the F.R.B.-M.I.T.-Penn model of the U.S. economy. *Amer. Econom. Rev.* **62**, 902–917.

Nelson, C. R. and G. W. Schwert (1977). Short term interest rates as predictors of inflation: On testing the hypothesis that the real rate of interest is constant. *Amer. Econom. Rev.* **67**, 478–486.

Newbold, P. and C. W. J. Granger (1974). Experience with forecasting univariate time series and the combination of forecasts. *J. Roy. Statist. Soc. A* **137**, 131–146.

Pagan, A. R. and G. W. Schwert (1990). Alternative models for conditional stock volatility. *J. Econometrics* **45**, 267–290.

Pesaran, M. H. (1974). On the general problem of model selection. *Rev. Econom. Stud.* **41**, 153–171.

Pesaran, M. H. and A. Timmermann (1992). A simple nonparametric test of predictive performance. *J. Business Econom. Statist.* **10**, 461–465.

Ramsey, J. B. (1969). Tests for specification errors in classical least-squares regression analysis. *J. Roy. Statist. Soc.* B **2**, 350–371.

Satchell, S. and A. Timmermann (1992). An assessment of the economic value of nonlinear foreign exchange rate forecasts. Financial Economics Discussion Paper FE-6/92, Birkbeck College, Cambridge University.

Schnader, M. H. and H. O. Stekler (1990). Evaluating predictions of change. *J. Business* **63**, 99–107.

Seillier-Moiseiwitsch, F. and A. P. Dawid (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88**, 355–359.

Shiller, R. J. (1979). The volatility of long term interest rates and expectations models of the term structure. *J. Politic. Econom.* **87**, 1190–1219.

Stekler, H. O. (1987). Who forecasts better? *J. Business Econom. Statist.* **5**, 155–158.

Stekler, H. O. (1994). Are economic forecasts valuable? *J. Forecasting* **13**, 495–505.

Theil, H. (1961). Economic Forecasts and Policy. North-Holland, Amsterdam.

Weiss, A. A. (1995). Estimating time series models using the relevant cost function. Manuscript, Department of Economics, University of Southern California.

Weiss, A. A. and A. P. Andersen (1984). Estimating forecasting models using the relevant forecast evaluation criterion. *J. Roy. Statist. Soc. A* **137**, 484–487.

West, K. D. (1994). Asymptotic inference about predictive ability. Manuscript, Department of Economics, University of Wisconsin.

West, K. D., H. J. Edison and D. Cho (1993). A utility-based comparison of some models of exchange rate volatility. *J. Internat. Econom.* **35**, 23–45.

Winkler, R. L. and S. Makridakis (1983). The combination of forecasts. *J. Roy. Statist. Soc. A* **146**, 150–157.