MANAGEMENT SCIENCE
Vol. 36, No. 9, September 1990
Printed in U.S.A.

COMBINING FORECASTS: OPERATIONAL ADJUSTMENTS TO THEORETICALLY OPTIMAL RULES*

DAVID C. SCHMITTLEIN, JINHO KIM AND DONALD G. MORRISON
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104
Korea Air Force Academy, Department of Management and Economics, Sangsuri Namilmyon Chungwongun, Chungbuk 363-849, *Korea*
Anderson Graduate School of Management, University of California, Los Angeles, California 90024-1481

Clemen and Winkler (1985) have described the theoretical effectiveness of Winkler's (1981) formula for optimally combining forecasts. The optimality of Winkler's formula is, however, contingent on actually knowing the forecasters' statistical properties, i.e., the variances and covariances of their forecasts. In realistic applications, of course, these properties have to be estimated, usually from a set of prior forecasts. In this case we show how the "operationally optimal" combining strategy differs from Winkler's "theoretically optimal" formula. Specifically, we provide figures indicating the operationally optimal strategy for combining two forecasts. We then propose a heuristic to choose the best set of parameter estimates in combining any number of forecasters and demonstrate its effectiveness via simulation.
(COMBINED FORECASTS; AKAIKE'S INFORMATION CRITERION)

1. Introduction

The challenge of finding a good combined forecast constructed from several separate forecasts has received a great deal of attention in this journal (Clemen 1987, Kang 1986, Gupta and Wilton 1987, Makridakis and Winkler 1983, Morris 1977, Winkler 1981) and elsewhere (Agnew 1985, Ashton 1986, Clemen and Winkler 1985, 1986, Geisser 1965, Newbold and Granger 1974). Makridakis and Winkler (1983) have discussed the practical interest in such combined forecasts, and Agnew (1985) has described the major methodologies proposed to date. Thus, we need not repeat those contributions here. Suffice it to say that combined forecasts have been useful in predicting economic time series data such as sales or GNP (Makridakis and Winkler 1983) and the market share of new products (Silk and Urban 1978).

This paper is concerned with predictions of a univariate quantity made by each of several forecasters. Winkler (1981) has shown how the optimal combination of the individual forecasts depends on the statistical properties of the forecasters (i.e., variances, covariances), when those properties are known. The resulting formula for combining the individual forecasts is a very convenient method for integrating information from dependent sources (i.e., correlated forecasts). The objectives of this paper are twofold. First, we illustrate the sensitivity of the combined forecast to the accuracy of, and correlations among, the individual forecasts. Second, we propose a heuristic to choose the parameter estimates in combining any number of forecasters. With regard to the first objectives, we plot in §2 the relative accuracy of the combined forecast as a function of these parameters that determine it. Although Bunn (1985) dealt with the same issue and presented series of tables showing relative performances of the six combining methods in simulations and case studies, the plots in this paper clearly show areas where each method outperforms others as a function of relevant parameters. Moreover, this paper seeks a new method to deal with cases of more than two forecasters. In §3 we propose a heuristic based on Akaike's Information Criterion (AIC) in choosing the best set of

* Accepted by Vijay Mahajan; received June 12, 1989. This paper has been with the authors 1 month for 1 revision.

0025-1909/90/3609/1044$01.25

parameter estimates to use and illustrate its effectiveness via simulation. The paper concludes with a discussion in §4 of some research issues that remain. There, we also show how our heuristic can be useful in combining predictions from uncalibrated (i.e., biased) forecasters as well.

2. Constructing the Operationally Optimal Combination of Two Forecasters

Imagine that assumptions (i) (multinormally distributed forecasts) and (ii) (calibrated forecasters) hold for a set of K forecasters. Each forecaster attempts to predict the same scalar quantity θ. Further, assume that a decision maker, who is not one of the forecasters, has a diffuse prior distribution regarding the forecast quantity θ. Then Winkler (1981) showed that the posterior distribution of the forecast quantity is normal with mean and variance

$$E[\theta \mid X, \Sigma] = e'\Sigma^{-1}X/e'\Sigma^{-1}e \qquad \text{and} \tag{1}$$

$$\text{Var}\,[\theta \mid X, \Sigma] = (e'\Sigma^{-1}e)^{-1} \tag{2}$$

respectively, where X is the K-vector of forecasts, Σ is the $K \times K$ covariance matrix of the errors in forecasts and $e' = (1, \ldots, 1)$ is the unit K-vector. In this Bayesian context equation (1) provides the forecast of θ, and (2) measures the accuracy of the combined forecast. Newbold and Granger (1974) have also derived equation (1), showing that it minimizes the squared forecast error.

In real forecasting settings properties of the forecasts are not known. Rather, a typical strategy would involve estimating them from past forecasts made by these forecasters. As several authors (e.g., Kang 1986) have observed, when the statistical properties of the forecasters must be estimated, the analyst can actually be better off by *ignoring* those estimates and assuming that a simpler process is operating; e.g., that the forecasters are uncorrelated; and/or that their accuracies are all equal. In the extreme case, it is possible that a simple unweighted average of the forecasts (which is optimal in theory if forecasters are equally accurate and uncorrelated) can outperform more complex combined forecasts based on estimated accuracies and correlations. Thus it is natural to wonder about constructing operationally optimal combined forecasts when a certain set of estimated properties have been obtained from a particular amount of information (i.e., previous forecasts).

In pursuing operationally optimal forecasts, we would like to retain the simplicity of Winkler's (1981) general formula. However, as above it is sometimes better *not* to substitute *all* of one's estimated forecast properties into the formula. The question for the analyst then becomes "which estimates do I use, and which ones should I ignore?" In this section we answer this question when two forecasts are being combined. In looking at the impact of the forecasters' correlations and accuracies on the combined forecast it is helpful to consider in detail the combination of just two forecasters, where the number of parameters is small enough to indicate their individual impacts on forecast accuracy.

Imagine that a complete set of forecasts (i.e., one from each forecaster) is available for each of M previous forecasting occasions. We will assume that the true statistical properties (ρ, σ_1, σ_2) of the forecasters are stationary both over the M previous forecasts, and over the future forecast(s) of interest.[1] The maximum likelihood estimates ($\hat{\rho}$, $\hat{\sigma}_1$, $\hat{\sigma}_2$) are easy to compute using the M previous observations, and can be substituted in (1) to combine future forecasts.

As mentioned above, however, one is sometimes better off *not* using all of the parameter estimates in formula (1). The added estimation error from using an additional estimated

[1] In this section, ρ denotes the correlation between the two forecasters, and σ_i the standard deviation of the forecast for forecaster i.

parameter can be greater than the reduction in modelling (misspecification) error associated with the additional parameter, as was demonstrated by Bunn (1985). (For a general discussion of this issue see Inagaki 1977.) In combining two forecasts there are three relevant parameters: the accuracy of each forecaster σ_1 and σ_2, and the correlation ρ. "Not using" the estimate $\hat{\sigma}_1$, $\hat{\sigma}_2$, and/or $\hat{\rho}$ in (1) would mean setting $\sigma_1 = \sigma_2$ and/or $\rho = 0$ in (1), respectively.

Choice of Models

Thus, in applying (1) the analyst has four model choices:

Model I. (Equal weights) Use values $\sigma_1 = \sigma_2$, $\rho = 0$ in (1) regardless of $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\rho}$.

Model II. Use values $\sigma_1 = \hat{\sigma}_1$, $\sigma_2 = \hat{\sigma}_2$, and $\rho = 0$ in (1).

Model III. Use values $\sigma_1 = \sigma_2$, $\rho = \hat{\rho}$ in (1).

Model IV. Use values $\sigma_1 = \hat{\sigma}_1$, $\sigma_2 = \hat{\sigma}_2$, and $\rho = \hat{\rho}$ in (1).

For combining two forecasts, however, Model III turns out to be equivalent to Model I (equal weights). This is easily seen by substituting ($\sigma_1 = \sigma_2$, $\rho = \hat{\rho}$) in (1) and noting that equal weights emerge. When the two forecasters are assumed to have equal accuracy ($\sigma_1 = \sigma_2$), the value of the correlation between them does indeed affect the *accuracy* of the optimal combined forecast, but it does not affect the *formula* for constructing that optimal combination.

Therefore, the purpose of this section is to see which of Models I, II and IV can be expected to perform better, as a function of the four relevant parameters (σ_1, σ_2, ρ, M). For these three models, a closed form expression for the expected accuracy of future combined forecasts (i.e., mean squared error) is not easy to obtain. Instead, an extensive simulation was conducted to reveal the relationship between this mean squared error of the combined forecast (MSE) and the parameters (σ_1, σ_2, ρ, M), for each of the three models.

The Simulation

In this simulation both forecasters were assumed to be calibrated, i.e., unbiased. Also, we arbitrarily set the variance of the first forecaster to 1, which enables us to examine the performance of the three models in a two-dimensional ($\sigma_2 \equiv \sigma$, ρ) space. For any particular (σ, ρ, M)-value studied, each replication involved randomly generating $M + 50$ observations from the bivariate normal distribution with mean vector (0, 0) and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}.$$

The first M observations are used to obtain maximum likelihood estimates ($\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\rho}$), which specify Models I, II and IV. Each model is then used to construct 50 combined forecasts, one for each of the 50 remaining observations. The mean squared error (MSE) is computed, for each model, across these 50 forecasts. Finally, this procedure is replicated 100 times to estimate the average MSE for each particular model and (σ, ρ, M)-value. Across the 100 replicates, the percentage of times that each model yielded the lowest MSE was also recorded, and is referred to as Percent Dominance below.

The average MSE and Percent Dominance, for each model, were computed for each of 1444 different (σ, ρ, M)-values, representing a full factorial design of:

M = 10, 25, 50 or 100 previous observations,

σ ranging from 1.0 to 2.9 in steps of 0.1, and

ρ ranging from −0.9 to 0.9 in steps of 0.1.

Across these 1444 design cells, we want to summarize which of Models I, II or IV should be selected by the analyst for making future combined forecasts.

The Results

The main results of the simulation study are shown in Figure 1. There, the model yielding the best average MSE is indicated, as a function of σ, ρ, and M. For example, consider an analyst having $M = 10$ previous forecasts with which to compute ($\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\rho}$) (i.e., the top left chart in Figure 1). If one forecaster is 40 percent more inaccurate than the other ($\sigma = 1.4$) and the correlation between forecasts is 0.4, then the operationally optimal strategy is to use Model II, i.e., use the observed estimate of $\hat{\sigma}_1$ and $\hat{\sigma}_2$, but ignore $\hat{\rho}$, setting $\rho = 0$. In the top left chart this is indicated by the point ($\sigma = 1.4$, $\rho = 0.4$) being in the region labeled for Model II. So even though the true correlation is nonzero in this example, the error introduced in trying to actually *estimate* ρ from 10 observations is greater than the value of the information gained in such a correlation.[2]

In general with $M = 25$ (top right chart in Figure 1), for equal weights (Model I) to be best, the forecaster's abilities must be similar ($\sigma < 1.1$) and large positive correlations must not occur ($\rho < 0.5$). Model II dominates when abilities are unequal ($\sigma > 1.2$) and the absolute value of the correlation is small ($-0.3 < \rho < 0.4$). It also dominates for similar abilities coupled with large positive correlations. Finally, Model IV dominates when abilities are unequal ($\sigma > 1.2$) and absolute correlations are large ($\rho < -0.3$, $\rho > 0.4$).

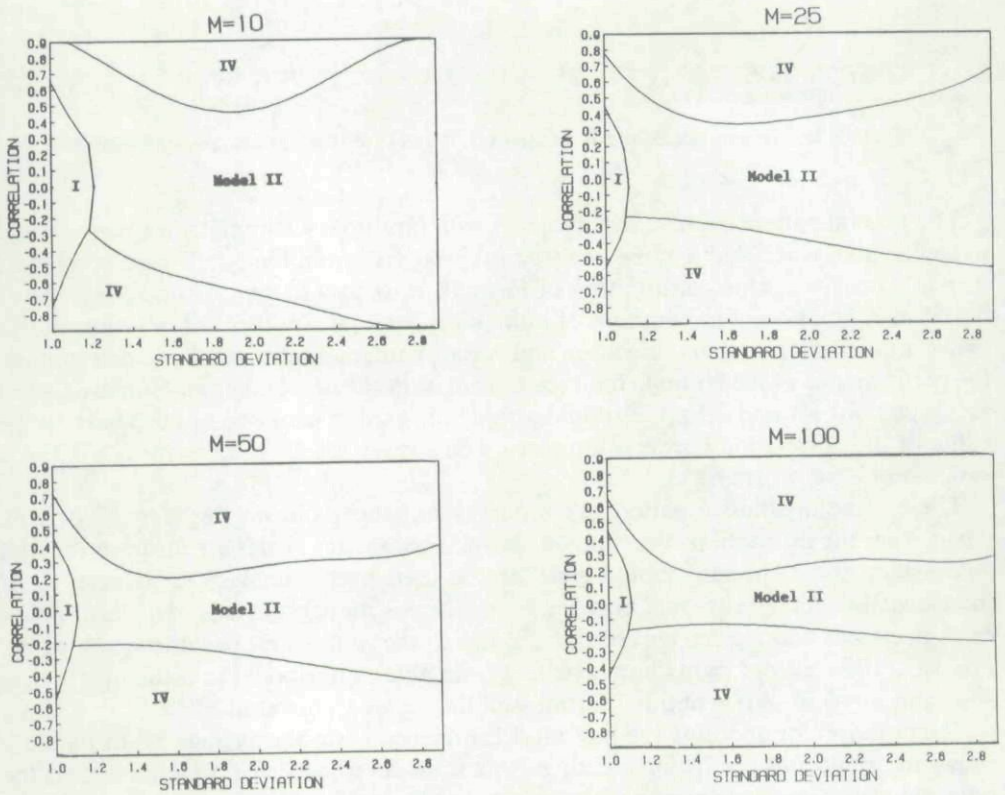

FIGURE 1. Performance Boundary (by Average MSE): Region in Which Each of Models I, II and IV is Dominant.

[2] The boundaries in Figures 1 and 2 have been smoothed to enhance interpretability, but are very close to the unsmoothed boundaries which were generated by the simulation and included in a previous draft of the manuscript. In fact, the smoothed and unsmoothed boundaries were so close that we were asked to delete the latter from the published version, and have complied.

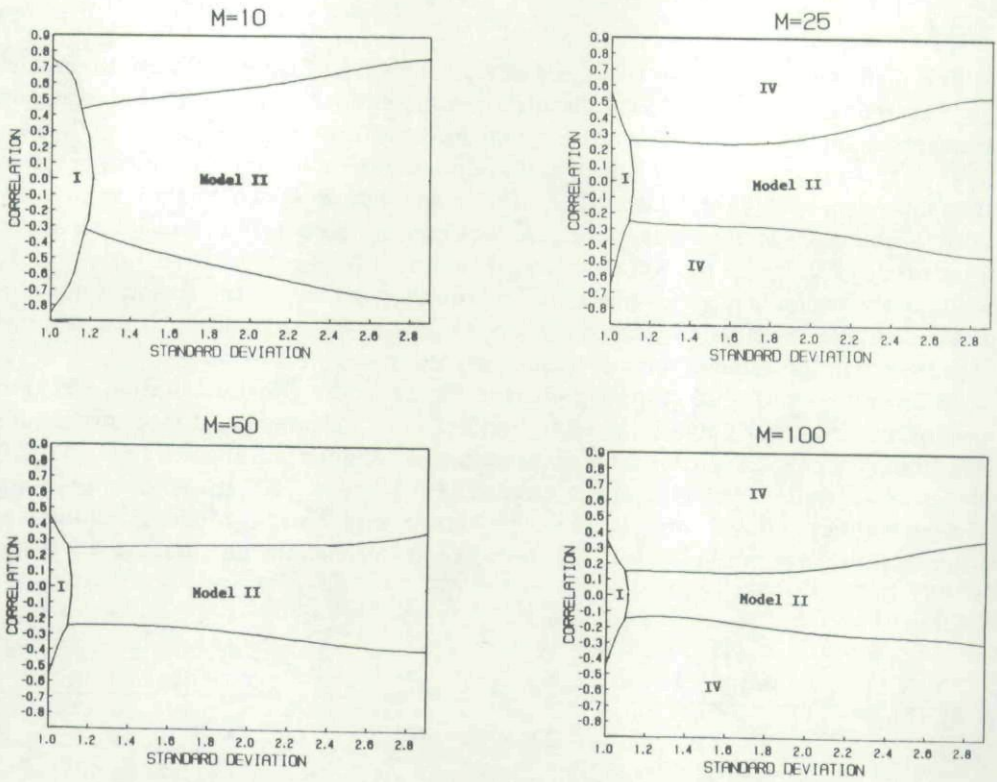

FIGURE 2. Performance Boundary of Models I, II and IV by Percent Dominance Criterion.

The general pattern of these results agrees with intuition: a parameter estimate should in fact be used when that parameter's true value is "far" from the "null" case $\sigma = 1$ (i.e., $\sigma_1 = \sigma_2$) or $\rho = 0$. One contribution of Figure 1 is to answer the question "How far is 'far'?" as a function of the amount of estimation data M. Another contribution in this figure is the revelation that correlation and variance interact substantially in determining the performance of the combined forecast. That is, deciding whether an estimated set of accuracies, $\hat{\sigma}_1 = 1$ and $\hat{\sigma}_2 = 1.2$, should actually be used in equation (1) depends on the value of the correlation between forecasts. The answer for $M = 25$ is yes if $-0.3 < \rho < 0.6$, and is no otherwise.

These same qualitative patterns of model dominance hold up for $M = 50$ or 100. Apart from the interaction effects noted above, if parameter values are far from the null values then their estimate should generally be used by the analyst. Of course, as M increases, the meaning of "far" changes. Since there is more estimation data, estimation error decreases and parameter estimates closer to the null values become worth using. For $M = 100$ (bottom right chart, Figure 1), we note that Model IV, using $\sigma_1 = \hat{\sigma}_1$, $\sigma_2 = \hat{\sigma}_2$ and $\rho = \hat{\rho}$ in (1), is optimal for most of the (σ, ρ)-values studied.

The criterion for choosing the best model in Figure 1 was the average MSE. Figure 2 shows the analogous results for selecting the best model when Percent Dominance is the criterion. The same general pattern emerges as in Figure 1, especially in deciding "How large is large enough?" for a parameter estimate to be useful in future predictions.

Figure 2's main difference from Figure 1 is that the boundary for selection of Model II versus Model IV shows more symmetry around the line $\rho = 0$. On reflection, this distinction also makes sense. The asymmetry in Figure 1 resulted mostly from a preference for Model II ($\sigma_1 = \hat{\sigma}_1$, $\sigma_2 = \hat{\sigma}_2$, $\rho = 0$) over Model IV ($\sigma_1 = \hat{\sigma}_1$, $\sigma_2 = \hat{\sigma}_2$, $\rho = \hat{\rho}$) when σ is small ($1 < \sigma < 1.2$) and ρ is large ($\rho > 0.5$). For these (σ, ρ) values Model IV is using

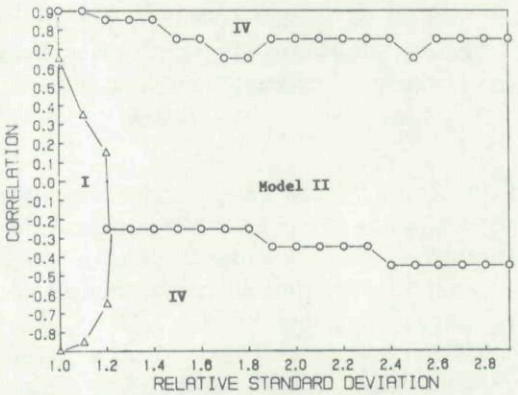

FIGURE 3. Performance Boundary Treating Only the Relative Accuracy and the Correlation as Unknown ($M = 10$, $\sigma_1 = 2$).

a "trick" to triangulate on the forecast quantity. That is, formula (1) actually assigns a negative weight to one of the two forecasts, and the other gets a large (>1) positive weight. The high positive correlation keeps the combined prediction around the true forecast quantity. But as Clemen and Winkler (1985) note, these negative weights, and the resulting forecasts, are very sensitive to small changes in the parameter values (σ, ρ) used.

By contrast, Model II plays it safe. Rather than using the "trick," this model just weights each forecast by its own accuracy. For these (low σ, high ρ) values, Figure 2 shows that Model IV's "trick" *usually* works best. But Figure 1 illustrates that when ($\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\rho}$) are relatively far from (σ, ρ). Model IV does very badly, giving it a higher average MSE over replications than Model II.

Recall that Figures 1 and 2 were created under the assumption $\sigma_1 = 1$. If the forecasters' parameters were known, making such an assumption would be no limitation at all, since equation (1) depends only on the *relative accuracy* (σ_r) of the second forecaster(i.e., relative to the first forecaster). But since the forecasters' parameters are estimated rather than known, this chosen value $\sigma_1 = 1$ could conceivably have had an effect on Figures 1 and 2 through its impact on the efficiency with which σ_r and ρ are estimated. This ought, however, to be a "second order" effect, and to demonstrate that Figures 1 and 2 are in fact insensitive to changes in σ_1, additional simulations were run, setting $\sigma_1 = 2$ and changing σ_r from 1 to 2.9. From a bivariate normal distribution with this covariance matrix and zero mean vector, we generated $M = 10$ observations which were used to estimate σ_r and ρ. The remaining steps of the simulation are the same as before. The results (unsmoothed) are summarized in Figure 3, which is very similar to the top left chart in Figure 1. This indicates that the results in Figures 1 and 2 are robust to our fixing σ_1, and treating only the relative accuracy and correlation as unknown.

Before turning to the case of $K > 2$ forecasters, we should mention one caveat regarding Figures 1 and 2. The data were simulated from a stationary process, where the forecasters' accuracies and correlation remain constant over the period examined. In real applications with some degree of nonstationary, presumably the relative performance of Model I (equal weights) would improve since it does not rely on a particular estimate of these nonstationary parameters.

3. Combining $K > 2$ Forecasts

We are again interested in deciding how many parameter estimates (and which ones) the analyst should actually use in constructing a combined forecast via equation (1). With $K > 2$ forecasters, the large number of parameters to estimate rules out an exhaustive

simulation study like the one above. Instead we will describe a simple heuristic based on Akaike's Information Criterion (AIC) for making this decision. Evidence will be presented that this heuristic is very effective in choosing the best set of parameter estimates to use.

Choosing Estimates Using AIC

For the purpose of combining forecasts, K calibrated forecasters are effectively characterized by their variances $(\sigma_1^2, \ldots, \sigma_K^2)$ and their correlations $(\rho_{ij}: i, j = 1, \ldots, K)$, i.e., by $K(K+1)/2$ parameters. Any particular decision of which parameters will take on the null value and which will be estimated can be thought of as proposing a model for describing the K forecasters. Since there are $L = K(K+1)/2$ parameters in all, each of which can be estimated or left as the null value, there are a total of 2^L such models.

Akaike (1974) proposed a simple criterion for choosing among models that captures the spirit of our concerns here. That is, a parameter estimate should only be used in a model if the reduction in misspecification error is greater than the error introduced in estimating this parameter. This interpretation of Akaike's Information Criterion has been explored by Inagaki (1977). Intuitively, such a criterion ought to work well in selecting a model that will perform well in making a variety of future (i.e., cross-validated) predictions. In fact, Stone (1974, 1977) showed that using AIC is asymptotically equivalent to choosing the model with the highest cross-validated likelihood value. Rust and Schmittlein (1985) have demonstrated that this asymptotic equivalence also holds up in small to moderate sample sizes for a variety of models. In the context of combining forecasts, Shibata (1980) used the AIC in selecting the optimal order for the autoregressive approximation and showed that the AR model, chosen by AIC, asymptotically attains the minimum mean squared prediction error. Diebold (1988) also relied on the AIC to specify an appropriate ARMA (p, q) model. Part of the appeal of Akaike's criterion is that it is so easy to compute. For a model based on T independently variable parameters $(\lambda_1, \ldots, \lambda_T)$ that are estimated from M observations $(x_1, \ldots, x_M)$, Akaike's Information Criterion is

$$A = -2(l(x_1, \ldots, x_M; \hat{\lambda}_1, \ldots, \hat{\lambda}_T) - T) \tag{3}$$

where $l(\ \ )$ is the log likelihood of the data and $(\hat{\lambda}_1, \ldots, \hat{\lambda}_T)$ are maximum likelihood estimates of the parameters. The model with the lowest value for A is to be preferred.

As (3) demonstrates, Akaike's criterion evaluates a model based on the maximum likelihood value that can be obtained, but then penalizes each model for the number of parameters estimated (T). In the results below, we will focus on a simple transformation of A that we will call AIC:

$$\text{AIC} = -\tfrac{1}{2}A = l(x_1, \ldots, x_M; \hat{\lambda}_1, \ldots, \hat{\lambda}_T) - T. \tag{4}$$

Criterion AIC is that actually considered by Stone (1979) and Rust and Schmittlein (1985). The model with the highest value of AIC is preferred. This criterion is easy to apply in our forecasting scenario when forecast errors are normally distributed. To evaluate AIC for any particular model (i.e., set of parameters that will be estimated for use in (1)), MLE's for the parameters are calculated, these MLE's are used to compute the likelihood of the M estimation observations, and the number of estimated parameters is subtracted from the likelihood.

How Well Does AIC Work?

Using AIC to choose the parameter estimates to actually use in (1) is easy for any number of forecasters, and it makes sense. But does it work? To answer this question, two separate simulations were conducted with two and five forecasters. Results from these simulations will be summarized in Figures 4 and 5, and Tables 1 and 2.

Using AIC To Combine Two Forecasters

Returning to the previous section's simulation study for two forecasters, the model yielding the largest average AIC value is plotted in Figure 4 as a function of the correlation $\hat{\rho}$ and relative accuracy $\hat{\sigma}$ of the forecasts. In this figure, $M = 10$ observations were again used in estimating σ_1, σ_2, and ρ. The similarity between Figure 4 and the top left chart in Figure 1 is striking. Using AIC to choose a model is very close to choosing the one most likely to yield the smallest mean squared error.

Figure 5 illustrates in more detail the case $\sigma = 2$. We have sketched both the average AIC and average MSE for Models II and IV as a function of the correlation ρ. For $M = 10$ estimation observations, charts (a) and (b) show the unsmoothed and smoothed functions, respectively. Similarly, charts (c) and (d) give the analogous results for $M = 25$.

In all four charts, the top two curves plot the average AIC values for Models II and IV. The bottom two curves sketch the average MSE for the same two models. For any value of the correlation, recall that the model having the *lowest* MSE is preferred, whereas the one having the *highest* AIC is selected. Figure 5 again highlights that choosing models based on AIC is very similar to choice based on the minimum average MSE. Although AIC and MSE are on different scales, the pattern of both sets of curves is very similar.

Combining Five Forecasters

In combining five forecasts, there are 15 relevant parameters. However, in applying (1) the analyst has four natural model choices as follows:

Model I:	Use values	$\rho_{ij} = 0$,	common $\sigma_i = \hat{\sigma}$ in (1),
Model II:	Use values	$\rho_{ij} = 0$,	$\sigma_i = \hat{\sigma}_i$ in (1),
Model III:	Use values	$\rho_{ij} = \hat{\rho}$,	common $\sigma_i = \hat{\sigma}$ in (1),
Model IV:	Use values common	$\rho_{ij} = \hat{\rho}$,	$\sigma_i = \hat{\sigma}_i$ in (1).

Model III turns out to be equivalent to Model I (equal weights). Therefore, we focus on Models I, II, and IV.

In this simulation each pair of forecasters was assumed to have the same underlying correlation ρ. The objective here is to investigate the effectiveness of the AIC criterion under a wide variety of reasonable parameter-value scenarios. Accordingly, the true underlying variance vector and common correlation for five forecasters were generated according to a gamma distribution (with mean 2 and variance 2) and a beta distribution (with varying parameters), respectively. From each such set of values we created the

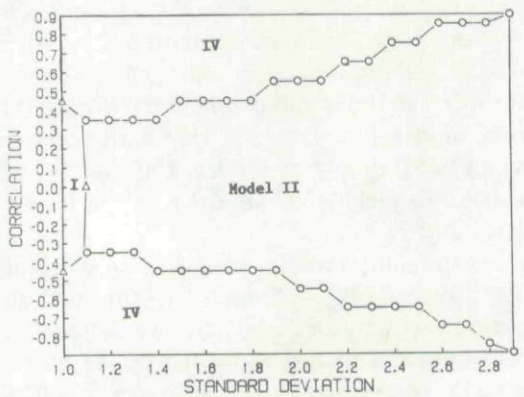

FIGURE 4. Performance Boundary of Models I, II and IV by AIC Criterion ($M = 10$, $K = 2$).

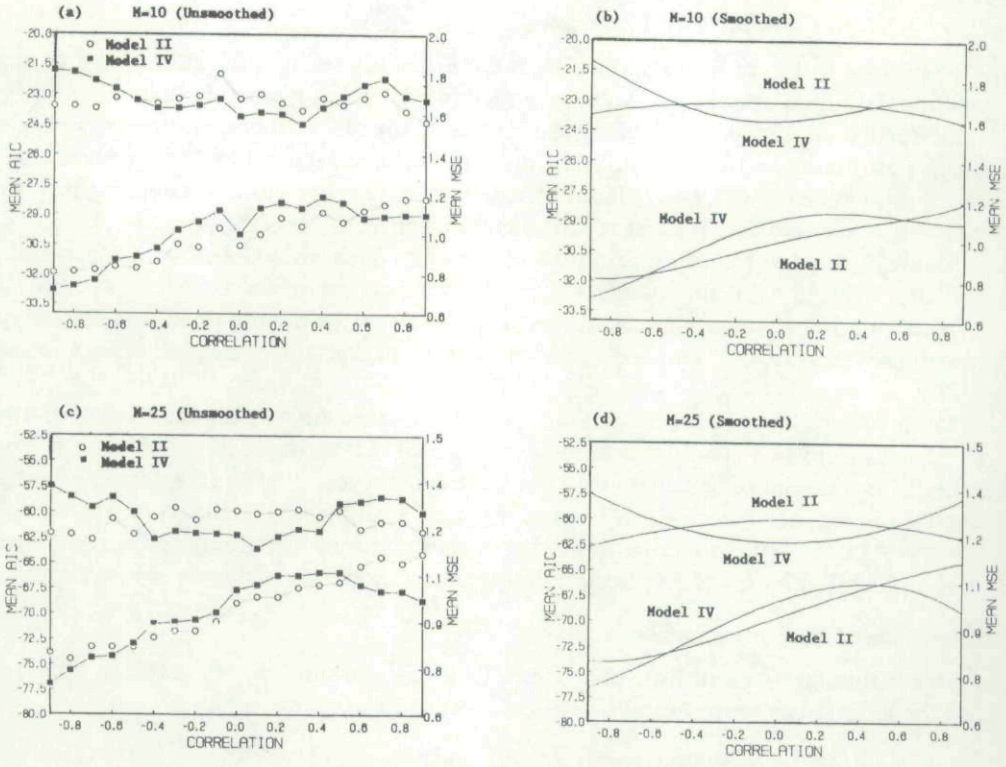

FIGURE 5. Comparison of Results from AIC and Average MSE ($\sigma = 2$).

corresponding population covariance matrix. Observations were then generated from the multinormal distribution with zero mean vector and this covariance matrix. The number of observations was determined by a truncated Poisson ($M \geq 6$) with varying means.

These observations were used to calibrate maximum likelihood estimates of the 10 pairwise correlations and 5 variances. A common correlation and variance were obtained by averaging these estimates. Then, to evaluate AIC for each of the three models, we computed the log likelihood of the observations and subtracted the number of estimated parameters from it. At this stage the model with the maximum AIC was selected. The purpose of this simulation is to see whether using AIC to choose a model works well. To investigate this, we generated a validation sample of 25 observations from the same population. Each model is then used to construct 25 combined forecasts corresponding to each of the observations. The MSEs are computed, for each model, across these 25 forecasts. Then, the MSE for the model picked by AIC was recorded separately. We replicated this procedure 1,000 times and obtained the average MSE for each of Models I, II and IV, and for the models picked by AIC. We anticipated that the latter should be smaller than the average MSE of any of Models I, II and IV, which indicates that the AIC works on the average in picking the model with the lowest MSE. The results are summarized in Tables 1 and 2.

First, Table 1 presents simulation results when $E[\rho] = 0.6$, $\text{Var}[\rho] = 0.04$, $E(M) = 6$ (i.e., Poisson-distributed with $E(M) = 6$, then left-truncated at $M = 6$), and each σ_i from a gamma distribution with mean 2 and variance 2. Panels (A) and (B) in Table 1 reveal that AIC picks the model with the lowest MSE in 66.1% of the cases (sum of diagonal elements in (B)). To evaluate the performance of AIC we use as a benchmark the average MSE for methods picked randomly, based on the marginal distribution of

TABLE 1

Performance of AIC Criterion ($E[\rho] = 0.6$, $E(M) = 6$ *and* $\sigma \sim$ *Gamma* (2, 1))

(A) Percentage that each model is picked

	Model I	Model II	Model IV
By AIC	72.6%	7.2%	20.2%
By MSE	74.2%	19.1%	6.7%

(*)

(B) Choice matrix (AIC by MSE, %)

By AIC \ By MSE	Model I	Model II	Model IV
Model I	60.0	10.9	1.7
Model II	3.1	2.6	1.5
Model IV	11.1	5.6	3.5

(C) Average MSE for models picked randomly by (*) = 0.7153

(D) Average MSE

	Model I	Model II	Model IV
Average MSE	0.6368	0.8405	1.2300
% Improvement if AIC used	−4%	21%	46%

Average MSE for methods picked by AIC = 0.6652.

each model chosen by MSE (see panel (A) in Table 1). We observe that, by comparing panels (C) and (D), AIC outperforms the benchmark model.

The results in Figures 4 and 5 above (for two forecasters) and these results in Table 1 indicate that AIC is an effective and useful criterion in deciding which estimated forecaster properties to use in combining forecasts. Of course, the results for combining five forecasts are averages over a wide range of specific forecast scenarios (i.e., true parameter values operating), which is necessary in view of the relatively large number of (variances, covariances) involved.

One final point deserves mention regarding Table 1. Note from panel (D) that, had one simply chosen Model I for all scenarios faced in this simulation, on average one would have done (very slightly) better than with AIC as the model selection criterion. Of course, such a comparison is somewhat unfair to AIC, since choosing Model I regardless of the estimates obtained is essentially "deciding not to decide" on which parameters to use. Nevertheless, it does reinforce robustness of equal weights (i.e., Model I) that has been found in other studies. Naturally, Model I will not perform the best for all reasonable simulation scenarios, and having *some* criterion for deciding which estimates to use remains important. AIC is seen to fulfill that role.

Table 2 shows the performance of AIC in various simulation settings. We used a full factorial $3 \times 3 \times 3$ design, jointly varying the distribution of σ_i, expected correlation $E[\rho]$, and average number of past observations for parameter estimation $E(M)$. The results show the same pattern as the findings in Table 1. That is, they indicate that the performance of AIC is not sensitive to the particular simulation scenario chosen for analysis. The first figure in each cell denotes how well AIC works compared to the method

TABLE 2

Summary of AIC Performance as a Function of Simulation Parameters

		Distribution for σ_i		
$E(\rho)$	$E(M)$	Mean = 2, Var = 2	Mean = 1.5, Var = 1.5	Mean = 1, Var = 1
$E(\rho) = 0.3$	$E(M) = 3$	(0.97[a], 0.97[b])	(0.97, 0.96)	(0.95, 0.96)
	$E(M) = 6$	(0.97, 0.98)	(0.97, 0.95)	(0.97, 0.93)
	$E(M) = 11$	(0.98, 0.95)	(0.97, 0.95)	(0.97, 0.95)
$E(\rho) = 0.6$	$E(M) = 3$	(0.95, 0.93)	(0.96, 0.92)	(0.94, 0.85)
	$E(M) = 6$	(0.96, 0.92)	(0.97, 0.92)	(0.93, 0.85)
	$E(M) = 11$	(0.96, 0.93)	(0.95, 0.89)	(0.95, 0.86)
$E(\rho) = 0.8$	$E(M) = 3$	(0.94, 0.86)	(0.95, 0.87)	(0.93, 0.90)
	$E(M) = 6$	(0.95, 0.86)	(0.97, 0.85)	(0.95, 0.91)
	$E(M) = 11$	(0.96, 0.85)	(0.97, 0.88)	(0.98, 0.89)

$$a = \frac{\text{Minimum average MSE among three models}}{\text{Average MSE for models picked by AIC}},$$

$$b = \frac{\text{Average MSE for models picked by AIC}}{\text{Average MSE for models picked randomly}}.$$

with the lowest MSE. These figures overall are very close to 1, which indicates that the heuristic performs reasonably well, relative to the (*post hoc*) best method. The second figure denotes the performance of AIC versus the benchmark described above. In all cases AIC outperforms this "naive" model. Notice that as the mean correlation ($E(\rho)$) increases, the gap in performance between AIC and the naive model widens.

4. Discussion

In this paper we have tried to describe some of the theoretical and practical properties of combined forecasts. On the theoretical side, they indicate the role that variances and correlations play in determining the best combined forecast. On the practical side, our results provide some guidance to a person who has multiple forecasts available, and who has an estimate of the statistical properties (variances, correlations) of those forecasts. In particular, we have shown which of these parameter estimates will actually improve the combined future forecasts if they are used in (1). Explicit guidelines are given for combining two forecasts, and a heuristic based on Akaike's Information Criterion is proposed for $K > 2$ forecasts. Finally, we presented some evidence that the AIC heuristic works well.

This research could be extended in several directions, including consideration of alternative distributional assumption (nonnormal), forecast combination formulas (nonlinear), and forecast evaluation criteria (besides MSE). To save space, we will briefly consider just two potential extensions that seem most interesting.

Uncalibrated Forecasters

Like many previous studies in this area, we have assumed that each forecaster is calibrated, i.e., unbiased. If instead some forecasters were known to be biased, and that bias

were also known, then formula (1) is easily modified to obtain the optimal combined forecast. (The bias is simply subtracted from each forecast before applying (1).)

In real applications, of course, these biases must be estimated from past forecasts along with the variances and correlations. As with the latter two quantities it is reasonable to ask whether the analyst is better off estimating these biases and using the estimates in (1) or instead assuming that the biases are zero (i.e., foregoing the estimation error).

The AIC criterion can resolve this issue, in the same way that it was used in the last section. That is, compute the maximized log-likelihood of the M estimation observations with, and without, the bias parameters being estimated. If the increase in the log-likelihood from the latter to the former case exceeds the number of extra (bias) parameters estimated, then using those estimations should indeed improve future predictions. That is, it should then pay off to attempt to calibrate the forecasters.

Shrinkage Estimates

In this paper we assume that the analyst either uses a parameter estimate "as is" in (1), or does not use it at all (i.e., sets it to the "null value"). As a hybrid or in-between strategy, one can consider "shrinking" the observed parameter estimate part of the way toward the null value. Presumably the amount of shrinkage would be inversely related to one's confidence in the observed estimate. Gupta and Wilton (1987), for example, adopt this general perspective.

Although appealing as a principle, there are some unresolved challenges in its application. Especially problematic is determining the amount of shrinkage to use for each of the parameters, and in particular whether that amount should be determined independently for each parameter (i.e., without looking at the other estimates). Relying on subjective or *ad hoc* procedures to set shrinkage levels is possible, but not very reassuring.

There is one way to avoid this dilemma and arrive at model-based shrinkage estimates. Namely, one can view the K forecasters as a random sample of size K from a superpopulation of forecasters. The estimated heterogeneity in bias and in accuracy across forecasters can be used to determine the appropriate shrinkage level for these two quantities (bias, accuracy) using standard empirical Bayes methods (Morris 1983). But a major problem remains. Adopting the "superpopulation" of forecasters notion means viewing the K forecasters at hand as a *random sample.* Thus, idiosyncratic patterns of *correlation among* these forecasters goes unanalyzed. Now, if there is one consistent finding from empirical studies of multiple forecasts, it is that the forecasts are *not* uncorrelated. Thus, a major remaining challenge for empirical Bayes shrinkage methods is the incorporation of contemporaneous correlation among the forecasts.

In general, we hope that the work in this paper has given the reader some insights into the nice theoretical results obtained by Winkler (1981) and Clemen and Winkler (1985). We hope that it stimulates further research on the properties of these combined forecasts in realistic application contexts. On the other hand, it will be difficult to beat "equal weight" forecasts in most practical situations. Even our $M = 10$ scenario is optimistic. Few important forecasting situations have as many as 10 previous forecasts from the same set of forecasters where the period of observation can be viewed as stationary. For example, with two forecasters and 10 good previous observations one needs fairly severe differences in ability and/or large correlations for something other than equal weights to be operationally optimal.

References

AGNEW, C., "Multiple Probability Assessments by Dependent Experts," *J. Amer. Statist. Assoc.*, 80 (1985), 343–347.

AKAIKE, H., "New Look at the Statistical Model Identification," *IEEE Trans. Automatic Control*, 19 (1974), 716–723.

ASHTON, R. H., "Combining the Judgments of Experts: How Many and Which Ones?" *Organizational Behavior and Human Decision Processes*, 38 (1986), 405–414.

BUNN, D. W., "Statistical Efficiency in the Linear Combination of Forecasts," *Internat. J. Forecasting*, 1 (1985), 151–163.

CLEMEN, R. T., "Combining Overlapping Information," *Management Sci.*, 33 (1987), 373–380.

——— AND R. L. WINKLER, "Limits for the Precision and Value of Information from Dependent Sources," *Oper. Res.*, 33 (1985), 427–442.

DIEBOLD, F. X., "Serial Correlation and the Combination of Forecasts," *J. Business and Economic Statist.*, 6, 1 (1988), 105–111.

GEISSER, S., "A Bayes Approach for Combining Correlated Estimates," *J. Amer. Statist. Assoc.*, 60 (1965), 602–607.

GUPTA, S. AND P. WILTON, "Combination of Forecasts: An Extension," *Management Sci.*, 33 (1987), 356–372.

INAGAKI, H., "Two Errors in Statistical Model Fitting," *Ann. Inst. Statist. Math., Part A*, 29 (1977), 131–152.

KANG, H., "Unstable Weights in the Combination of Forecasts," *Management Sci.*, 32 (1986), 683–695.

MAKRIDAKIS, S. AND R. L. WINKLER, "Averages of Forecasts: Some Empirical Results," *Management Sci.*, 29 (1983), 987–996.

MORRIS, C. N., "Parametric Empirical Bayes Inference: Theory and Application," *J. Amer. Statist. Assoc.*, 78 (1983), 47–55.

MORRIS, P. A., "Combining Expert Judgments: A Bayesian Approach," *Management Sci.*, 23 (1977), 679–693.

NEWBOLD, P. AND C. W. J. GRANGER, "Experience with Forecasting Univariate Time Series and the Combination of Forecasts," *J. Roy. Statist. Soc. Ser. A*, 137, Part 2 (1974), 131–164 (with discussion).

RUST, R. T. AND D. C. SCHMITTLEIN, "A Bayesian Cross-Validated Likelihood Method for Comparing Alternative Specifications of Quantitative Models," *Marketing Sci.*, 4 (1985), 20–39.

SILK, A. J. AND G. L. URBAN, "Pre-Test Market Evaluation of New Packaged Goods: A Model and Measurement Methodology," *J. Marketing Res.*, 15 (1978), 171–191.

STONE, M., "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. Roy. Statist. Soc. Ser. B*, 36 (1974), 111–147.

———, "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *J. Roy. Statist. Soc. Ser. B*, 39 (1977), 44–47.

———, "Comments on Model Selection Criteria of Akaike and Schwarz," *J. Roy. Statist. Soc. Ser. B*, 41 (1979), 276–278.

WINKLER, R. L., "Combining Probability Distributions from Dependent Information Sources," *Management Sci.*, 27 (1981), 479–488.