7
Combining Forecasts with Many Predictors

Tae-Hwy Lee
University of California, Riverside

In practice it is quite common that one forecast model performs well in certain periods while other models perform better in other periods. It is difficult to find a forecast model that outperforms all competing models. To improve forecasts over individual models, combined forecasts have been suggested (Bates and Granger 1969). Researchers including Newbold and Granger (1974), Granger and Newbold (1986, Ch. 9), Granger and Jeon (2004), and Yang (2004) show that forecast combinations can improve forecast accuracy over a single model and show why the forecast combination can achieve a better forecast in terms of mean squared forecast error. Bayesian model averaging may be used to form a weighted combined forecast. (See, e.g., Lee and Yang [2006].) A matter frequently discussed in the literature is how to combine forecasts to achieve the most accurate result. (See Granger and Ramanathan [1984]; Deutsch, Granger, and Teräsvirta [1994]; Palm and Zellner [1992]; Shen and Huang [2006]; and Hansen [2008].) Clemen (1989) and Timmermann (2006) provide excellent surveys on forecast combination and related issues.

Granger and Jeon (2004, p. 327) put the forecast combination in a general context of *thick* modeling and write, "An advantage of thick modeling is that one no longer needs to worry about difficult decisions between close alternatives or between deciding the outcome of a test that is not decisive. In time series such questions are whether the process has a unit root or not, or how many cointegrations are in a vector of a series. For thick models one considers all plausible alternatives and uses the outputs of the various models."

Even when we have a single model, a combination of forecasts can also be formed over a set of training sets. While, in practice, usually

we have a single training set, it can be replicated via bootstrap, and the combined forecast trained over the bootstrap-replicated training sets can improve upon the original forecast of the model. This is the idea of bootstrap aggregating (abbreviated as "bagging"), introduced by Breiman (1996).

Huang and Lee (2010) consider the situation in which one wants to predict an economic variable using the information set of many relevant explanatory variables. As Diebold and Pauly (1990, p. 503) point out, "It must be recognized that in many forecasting situations, particularly in real time, pooling of information sets is either impossible or prohibitively costly." Likewise, when models underlying the forecasts remain partially or completely unknown (as is usually the case in practice—for example, with survey forecasts), one would never be informed about the entire information set. Quite often the combination of forecasts is used when the only things available are individual forecasts (for example, in the case of professional forecasters), while the underlying information set and the model used for generating each individual forecast are unknown.

In this chapter we consider how to combine forecasts in a situation where *many* predictors (in other words, a large information set) are available, or in a situation where *many* forecasts are given but models and predictors used for generating each individual forecast are not necessarily known. In each of these situations, we explain how to use factor models. Much of the results presented here are studied in Chan, Stock, and Watson (1999); Hillebrand et al. (2010); Huang and Lee (2010); Stock and Watson (2002); and Tu and Lee (2009).

DATA-RICH ENVIRONMENT

Bernanke and Boivin (2003) emphasize that the use of a large data set is a common practice, such as in the central bank's policymaking analysis. They write, "Research departments throughout the Federal Reserve System monitor and analyze literally thousands of data series from disparate sources . . . Despite this reality of central bank practice, most empirical analyses have been confined to . . . exploit only a limited amount of information. For example, the VAR methodology generally

limits the analysis to eight macroeconomic time series or fewer. This disconnect between central bank practice and academic analysis has several costs . . . It thus seems worthwhile to take into account the fact that in practice monetary policy is made in a data-rich environment" (p. 526).

For example, in forecasting stock market volatility, we can use many predictors from many options' implied volatilities. In predicting output growth and inflation, we can use many available economic predictors (Bernanke and Boivin 2003; Hillebrand et al. 2010; Stock and Watson 2002; Tu and Lee 2009; Wright 2009). Ang and Piazzesi (2003); Ang, Piazzesi, and Wei (2006); Bernanke (1990); Hillebrand et al. (2010); and Stock and Watson (1989) use many yields and yield spreads. To predict retail default probability, a retail credit model uses many borrower-specific predictors.

Bernanke and Boivin (2003, p. 525) confirm the merit of the large data set: "[It] explores the feasibility of incorporating richer information sets into the analysis, both positive and normative, of Fed policy making. We employ a factor-model approach . . . that permits the systematic information in large data sets to be summarized by relatively few estimated factors. With this framework, we confirm Stock and Watson's result that the use of large data sets can improve forecast accuracy . . ."

A natural question arises as to how we should use all those vast data in predicting a target of interest. Using large data, there are advantages to accessing rich information and robustifying against structural instability, which plagues low-dimensional forecasting. While we can exploit these advantages, there are also difficulties attached to using large data due to overwhelming information, which may be highly correlated and noisy.

When there are many predictors in columns of the predictor matrix X with the column number N being large, the dimension N needs to be reduced. One way is to select r ($\ll N$) factors of X, and another way is to select r ($\ll N$) columns of X. The former way, known as a factor model, has recently been a popular approach in economic forecasting, because of pioneering work by Stock and Watson (2002), Bai (2003), and Bai and Ng (2002, 2006), who have explored theoretical and empirical analysis of factor models based on principal components. The latter way, known as variable selection, has been widely studied in statistics. The variable selection serves to reduce N by ranking and selecting a

subset of X that is most predictive for a forecast target y, through such methods as LASSO (least absolute shrinkage and selection operator) (Tibshirani 1996), least angle regression (Efron et al. 2004), and elastic net (Zou and Hastie 2005), among many other methods.

While the data-rich environment usually refers to the situation where there are many predictors, it also refers to the situation where there are many forecasts provided by many firms, many departments in an organization, many analyses in an investment bank, many different government agents, and so on. In this paper we consider both cases—namely, the data-rich environment with many predictors with N-vector $\boldsymbol{x}_t = (x_{t1}, x_{t2},\ldots,x_{tN})'$, and the data-rich environment with many forecasts with N-vector of forecasts $\hat{\boldsymbol{y}}_{t+h} = \left(\hat{y}^{(1)}_{t+h}, \quad \hat{y}^{(2)}_{t+h}, \quad \ldots, \quad \hat{y}^{(N)}_{t+h}\right)'$. Below we discuss how we form a forecast under these two types of data-rich environment. In both cases the idea is to combine multiple forecasts. Therefore we begin with a review of the literature on combining forecasts.

When multiple forecasts of the same variables are available, it's typically argued that a combination of those forecasts should be used instead of using any single forecast, even if it's a dominant one (e.g., Timmermann 2006). This is because forecast combinations offer diversification gains, and it's almost impossible to identify ex ante a dominant forecast model. The success of the forecast combinations will in turn depend on how well the combination weights are determined. As summarized in Clemen (1989), a simple average (with weights $\frac{1}{N}$) of the multiple forecasts is typically found to be a good forecast combination. However, the equal weights $\frac{1}{N}$ will be very small when N is very large in a data-rich environment, giving little chance for a better model to work dominantly against bad models. Before we deal with the data-rich environment, we first consider a simplest-case scenario, that of $N = 2$.

COMBINING FORECASTS

Bates and Granger (1969) first introduced the idea of combining forecasts. Let us begin with their brief review of what happens when $N = 2$. Let $\hat{y}^{(1)}_t$ and $\hat{y}^{(2)}_t$ be forecasts of y_{t+1} with errors

$e_{t+1}^{(i)} = y_{t+1} - \hat{y}_t^{(1)}, \qquad i = 1,2,$

so that $Ee_{t+1}^{(i)} = 0, \quad Ee_{t+1}Ee_{t+1}^{(i)2}{}_{i} = \sigma_i^2$, and $Ee_{t+1}^{(1)}e_{t+1}^{(2)} = \rho\sigma_1\sigma_2$.

Define a combined forecast with the weight $w \in (-\infty\ \infty)$,

$\hat{y}_t^{(c)} = w\hat{y}_t^{(1)} + (1-w)\hat{y}_t^{(2)},$

its forecast error

$e_{t+1}^{(c)} = y_{t+1} - \hat{y}_t^{(c)} = we_{t+1}^{(1)} + (1-w)e_{t+1}^{(2)},$

and its expected squared forecast error loss

$\sigma_c^2(w) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\rho\sigma_1\sigma_2.$

Minimizing the loss, the optimal combining forecast weight is obtained.

This expression is minimized for the value of k given by

$$(7.1) \quad w_{opt} = \arg\min \sigma_c^2(w) = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Substitution yields the minimum achievable error variance as

$$\sigma_c^2(w_{opt}) = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Bates and Granger (1969) show that the optimal combined forecast error loss is smaller than the smaller of the two individual forecast error losses:

$\sigma_c^2(w_{opt}) \le \min(\sigma_1^2, \sigma_2^2).$

Thus, a priori, it is reasonable to expect in most practical situations that the best available combined forecast will outperform the better individual forecast. It cannot, in any case, do worse.

This result has been used across various disciplines (e.g., economics, finance, operations research, meteorology, management, computer

science, and machine learning) under the names of combining forecast, ensemble predictor, committee of learners, team of forecasts, consensus of learners, mixture of experts, expert system, and others.

WHY COMBINE?

The forecast combination problem is similar to that of minimizing the variance of a portfolio, with the errors from the individual forecasts playing the role of asset returns (Aiolfi and Timmermann 2006). In practice it is quite common that one forecast model performs well in certain periods while other models perform better in other periods. It is difficult to find a forecast model that outperforms all competing models. Forecast combinations can improve forecast accuracy over a single model. Hong and Lee (2003) find that the combined forecasts are generally the best performer for the mean, and sign prediction for the foreign exchange rate changes.

Aiolfi and Timmermann (2006) consider a forecasting strategy that takes the average over the models in the top quartiles or cluster. There is clear evidence that, in general, a strategy of selecting one best (top) model based on past forecasting performance does not work well. This holds true both for linear and nonlinear forecasting methods. This is analogous to portfolio selection in the stock market.

Why do we combine? Aiolfi and Timmermann (2006) answer this way: "Forecast combination entails using information from a typically large set of forecasts and emerges as an attractive strategy when individual forecasting models are misspecified in a way that is unknown to the modeler. Misspecification is likely to be related not simply to functional form (neglected nonlinearity) but also to instability (structural changes) in the joint distribution of forecasts and the target variable. In this situation, the identity of the best forecasting model is likely to change over time and a key question is for how long the relative performance of forecasting models persists" (p. 33).

Aiolfi and Timmermann (2006, p. 32) also write the following:

> Forecasts are of considerable importance to decision makers throughout economics and finance and are routinely used by private enterprises, government institutions and professional econ-

> omists. It is therefore not surprising that considerable effort has gone into designing and estimating forecasting models ranging from simple, autoregressive specifications to complicated nonlinear models or models with time-varying parameters. A reason why such a wide range of forecasting models is often considered is that the true data generating process underlying a particular series of interest is unknown. Even the most complicated model is likely to be misspecified and can, at best, provide a reasonable "local" approximation to the target variable.

I would add that this is particularly so in practical forecasting situations in macroeconomics with a large cross section of forecasting models and a short time-series dimension. Aiolfi and Timmermann (2006, p. 32) go on to make a second point about forecasting models:

> Model instability is a source of misspecification that is likely to be particularly relevant in practice, c.f. Stock and Watson (1996). In its presence, it is highly unlikely that a single model will be dominant uniformly across time and the identity of the best local approximation is likely to change over time. If the identity of the best local model is time-varying, it is implausible that a forecasting strategy that, at each point in time, attempts to select the best current model will work well. Most obviously, if (ex-ante) the identity of the best model varies in a purely random way from period to period, it will not be possible to identify this model by considering past forecasting performance across models. Similarly, if a single best model exists but only outperforms other models by a margin that is small relative to random sampling variation, it becomes difficult to identify this model by means of statistical methods based on past performance. Even if the single best model could be identified in this situation, it is conceivable that diversification gains from combining across a set of forecasting models with similar performance will dominate the strategy of only using a single forecasting model.

HOW TO COMBINE?

The optimal combination weights in Equation (7.1) for $N = 2$ may be extended to a general case with a larger N. However, the estimation

of the weights from the regression of the form (Equation [7.2]) may suffer from a large estimation error, especially when N is large, and the forecasts may be highly correlated. The following methods have been widely used in applications.

A natural way is to estimate forecast combination weights by least squares regression or, equivalently, by using portfolio variance minimization methods. The usual problem with this estimation method is that, given the sample sizes typically available in practice, the combination weights are often imprecisely estimated. In particular, this is a problem when the number of models is large relative to the length of the time series, so that the covariance matrix of the forecast errors either cannot be estimated or is estimated very imprecisely. The assumption of a stable covariance structure is unlikely to be satisfied in practice, and weights may be time-varying.

A simpler way is to use the equal weights (simple mean). This becomes a common strategy when the models are of similar quality or when their relative performance is unknown or unstable over time. Stock and Watson (1999) use trimmed mean and median to robustify the simple mean–weighted combined forecasts.

Aiolfi and Timmermann (2006) use ranking of the forecasting model and also use clustering. The premise of this approach is that, when combining forecasts from a large cross section of models, it is generally difficult to distinguish between the performance of the top models, but one can tell the difference between the best and worst models. This suggests including a subset of "good" models in the combined forecast. Another popular method is Bayesian model averaging, which is used in many applications, for example, Lee and Yang (2006) and Wright (2009).

The formula for the optimal combination weights in Equation (7.1) for $N = 2$ has an important aspect that has been ignored in many applications in the literature, although it was discussed in Granger and Newbold (1986, Ch. 9) in some length and detail. That is the role of correlation ρ on the forecast combination as studied in Lee, Li, and Huang (2010). Note that the forecast combination need not be convex, and it is permitted that the weights can be any real number, $w \in (-\infty\ \infty)$. Therefore the optimal forecast combination weight w in Equation (7.1) may be negative (< 0) or larger than 1. What does this mean? How does ρ affect the combined forecast? To combine multiple forecasts when these

forecasts are highly correlated or close to collinear, the optimal combination places negative weights on the inferior forecasts and larger than 1 on the dominant forecasts, similar to the pairs-trading strategy that profits from the high correlation of the two sock returns. This optimal forecast combination outperforms any individual forecast and explains why an inferior forecast can be included in the combination to improve the forecast. The optimal combination weight has a pattern similar to that of the pairs-trading strategy. Without loss of generality, we assume all the forecasts are one-step-ahead forecasts. The following results can be easily generalized to multistep forecasts. The situation where $w_{opt} < 0$ is interesting. In light of the above condition, it appears that an inferior forecast may still be worth including with negative weight. This happens when $\sigma_2^2 - \rho\sigma_1\sigma_2 < 0$ or $\sigma_2 / \sigma_1 < \rho$—i.e., when ρ is a very large positive value (say, close to 1) and $f_t^{(1)}$ is the inferior forecast, with larger forecast error variance σ_1.

As shown in Granger and Newbold (1986, p. 268), the optimal combining weight w_{opt} can be estimated from

$$\text{(7.2)} \quad \hat{w}_t = \frac{\sum_{s=1}^{t}\left(e_s^{(2)2} - e_s^{(1)}e_s^{(2)}\right)}{\sum_{s=1}^{t}\left(e_s^{(1)2} + e_s^{(2)2} - 2e_s^{(1)}e_s^{2}\right)},$$

which can be obtained from the regression

$$\text{(7.3)} \quad e_{t+1}^{(2)} = w\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right) + e_{t+1}^{(c)}.$$

However, a common popular recommendation is to ignore ρ. For example, Clemen (1989, p. 562) suggests "to ignore the effect of correlations in calculating combining weights." While the optimal weight $\hat{w}_t$ can be negative or overweighted (larger than 1) depending on the value of ρ, the use of a simpler form obtained with the restriction $\rho = 0$ has been a popular recommendation:

$$\hat{w}_t' = \frac{\sum_{s=1}^{t} e_s^{(2)2}}{\sum_{s=1}^{t}\left(e_s^{(1)2} + e_s^{(2)2}\right)}.$$

Note that, if we ignore ρ, $\hat{w}_t'$ is always constrained on the (0,1) interval (analogous to the short-sale constraint).

When ρ is large and positive, the optimal weight on the inferior forecast can be negative. The forecast combination problem is analogous to that of minimizing the variance of a portfolio, with the forecast errors playing the role of asset returns (Timmermann 2006). Gatev, Goetzmann, and Rouwenhorst (2006) show that "pairs trading" in financial trading strategy profits from the high correlation in the returns. Analogously, the profitability of using the optimal weight is linked to the high correlation ρ in the forecasts. Without loss of generality, let us assume that $\hat{y}_t^{(1)}$ is the inferior forecast, with larger forecast error variance. In combining forecasts, when $\rho \gg 0$, we short the loser (the worse forecast) with $w < 0$ and buy the winner (the better one) with $(1 - w) > 1$. In this case, the use of $\hat{w}_t'$ while ignoring the correlation ρ would be too restrictive.

FORECASTING IN A DATA-RICH ENVIRONMENT

So far, we have looked at the case of $N = 2$. Most of the combining forecast literature has been limited to the case in which N is small. Now we take up the case of combining forecasts when N is large. Consider a kitchen-sink model with all predictors $\boldsymbol{x}_t$ in one large model

$$y_{t+h} = \left(1 \quad \boldsymbol{x}_t'\right)\boldsymbol{b} + u_t \qquad (t = 1, 2, \ldots, T)$$

to generate the h-step forecast

$$\hat{y}_{T+h} = \left(1 \quad \boldsymbol{x}_T'\right)\hat{\boldsymbol{b}}_T .$$

However, when N is large, the OLS estimator $\hat{b}_{OLS}$ may not be feasible to compute, and the mean squared forecast error (MSFE) increases with N as MSFE $= E\left(y_{t+h} - \hat{y}_{t+h}\right)^2 = O\left(\frac{N}{T}\right)$. A solution to these problems is not to use OLS estimation of the large model but to reduce the dimension N, either by selection of relevant variables for the forecast target to reduce N or by using a factor model to reduce N, or both. The variable selection is to reduce N by ranking variables in X and select-

ing a subset of X that is most predictive for a forecast target y, through such methods as forward and backward selection, stepwise regression, LASSO (Tibshirani 1996), least angle regression (Efron et al. 2004), elastic net (Zou and Hastie 2005), and so on.

Alternatively, one can combine the large information in $\boldsymbol{x}_t$ *indirectly* through individual forecasts $\hat{y}_{T+1}^{(i)}$ ($i = 1, \ldots, N$), and then combine the N individual forecasts

$$y_{t+1}^{(1)} = x_{1t}\beta_1 + \varepsilon_{1,t+1}$$

$$\vdots$$

$$y_{t+1}^{(N)} = x_{Nt}\beta_N + \varepsilon_{N,t+1}$$

to form the combined forecast (at time T using the estimated $\hat{\beta}_i$'s)

$$\hat{y}_{T+1}^{(c)} = w_1\hat{y}_{T+1}^{(1)} + \ldots + w_N\hat{y}_{T+1}^{(N)} .$$

Here each partition of the predictor vector $\boldsymbol{x}_t$ need not contain only one predictor at a time but may contain more, and each partition need not be disjointed. In practice, generally the predictor vector $\boldsymbol{x}_t$ may not be observed when only forecasts are available (e.g., in survey forecasts). Therefore, we will consider two types of data-rich environments. The first is where there are N predictors

$$\boldsymbol{x}_t = \left(x_{t1}, x_{t2}, \ldots, x_{tN}\right)' ,$$

and the second is where there are N forecasts, with

$$\hat{\boldsymbol{y}}_{t+h} = \left(\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \ldots, \hat{y}_{t+h}^{(N)}\right)' .$$

In each type of data-rich environment, we use factor models assuming there are latent factors of the predictors $\boldsymbol{x}_t$ or of the forecasts $\hat{\boldsymbol{y}}_{t+h}$.

FORECASTING WITH MANY PREDICTORS

First, let's consider forecasting when there are N predictors $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tN})'$ and N is large. Following Stock and Watson (2002), we use a factor model that is based on the factors f_t of the predictors $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t = \Lambda f_t + \nu_t \quad , \tag{7.4}$$

where Λ is the factor loading. Once the factors have been extracted from the predictors, the forecast of the target can be formed from the regression of

$$\boldsymbol{y}_t = f_t \alpha + u_t \quad . \tag{7.5}$$

As noted in Hillebrand et al. (2010) and Tu and Lee (2009), in this approach, the factors are obtained from the marginal model of $\boldsymbol{x}_t$ rather than the joint model of $(\boldsymbol{y}\ \boldsymbol{x}_t)$. We write the above model in Equations (7.4) and (7.5) as follows:

$$\boldsymbol{y}_t = E\left(y_t \left| \boldsymbol{x}_t; \theta_1\right.\right) + u_t = f_t \alpha + u_t \qquad \left(\theta_1 = \alpha\right),$$

$$\boldsymbol{x}_t = E(\boldsymbol{x}_t; \theta_2) + \upsilon_t = \Lambda f_t + \upsilon_t \qquad (\theta_2 = f_t, \Lambda) \ .$$

Note that this assures that the joint density

$$D\left(y_t, \boldsymbol{x}_t; \theta\right) = D_1\left(y_t \left| \boldsymbol{x}_t; \theta_1\right.\right) \times D_t\left(\boldsymbol{x}_t; \theta_2\right),$$

where $\left(\theta_1\ \theta_2\right) \in \Theta_1 \times \Theta_2$ are variation free and we estimate the conditional model (Equation [7.5]) and the marginal model (Equation [7.4]) separately.

FORECASTING WITH MANY FORECASTS

Next, we consider forecasting when there are N forecasts $\hat{\boldsymbol{y}}_{t+h} = \left(\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \ldots, \hat{y}_{t+h}^{(N)}\right)'$ and N is large. In this situation, many fore-

casts are given either from many survey forecasters or from many analysts. There are various organizations that operate as an aggregate or a group, based on many individual analysts who may or may not use the same information sets. Depending on the shared intersections of various information sets used by survey forecasters or analysts, the correlations among the many individual forecasts may be strong. When the number N of individual forecasts is large, we wish to estimate the weights to form the aggregate forecast (a combined forecast). The N individual forecasts may be given with or without the prescription on how they have been generated. We apply principal component analysis on the forecasts to extract factors

$$\hat{\boldsymbol{y}}_{t+h} = \Lambda f_{t+h} + \upsilon_{t+h}$$

and

$$\hat{f}_{t+h} = \hat{\Lambda}' \hat{\boldsymbol{y}}_{\boldsymbol{t+h}}$$

and estimate the following forecasting equation,

$$(7.6) \qquad y_{t+h} = \hat{f}'_{t+h}\alpha + u_{t+h} \ ,$$

to form the eventual forecast

$$\hat{y}_{T+h} = \hat{f}'_{T+h}\hat{\alpha}_T \ .$$

From the above calculations, note that the weights to combine many forecasts are

$$\hat{y}_{T+h} = \hat{f}'_{T+h}\hat{\alpha} = \left(\hat{y}'_{T+h}\hat{\Lambda}\right)\hat{a} = \hat{y}'_{T+h}\hat{w} \ ,$$

and therefore the optimal forecast combination weights are

$$\hat{w} := \hat{\Lambda}\hat{\alpha} \ .$$

Hillebrand et al. (2010) and Tu and Lee (2009) consider the above model when each individual forecast $\hat{y}^{(i)}_{t+h}$ is generated by using one predictor $x^{(i)}_t$ at a time. In their applications, the combined forecast with

this weight vector $\hat{w} = \hat{\Lambda}\hat{\alpha}$ outperforms the equally weighted combined forecast. However, it is not necessary to know how each individual forecast $\hat{y}_{t+h}^{(i)}$ is generated. In practice, there are various situations where only the forecasts are given to econometricians, without telling about how the forecasts are obtained.

It is generally believed that it is difficult to estimate the forecast combination weights when N is large. Therefore the equal weights $\left(\frac{1}{N}\right)$ have been widely used instead of estimating weights. An exception is Wright (2009), who uses Bayesian model averaging (BMA) for pseudo out-of-sample prediction of U.S. inflation and finds that it generally gives more accurate forecasts than simple equal-weighted averaging. He uses $N = 107$ predictors. It is often found in the literature that equally weighted combined forecasts are the best. Stock and Watson (2004) call this the "forecast combination puzzle." (See also Timmermann [2006].) Smith and Wallis (2009) explore a possible explanation of the forecast combination puzzle and conclude that it is due to estimation error of the combining weights. However, the empirical results occur when N is not very large. When N is very large, the equal weights $\left(\frac{1}{N}\right)$ put too little weight to good models, especially when $N \to \infty$, and the equal weights can hardly be justified. Note that we can consistently estimate the combining weights $\hat{w} = \hat{\Lambda}\hat{\alpha}$, as long as $\hat{\Lambda}$ and $\hat{\alpha}$ are estimated consistently. Note also that combining forecasts with the weights $\hat{w} = \hat{\Lambda}\hat{\alpha}$ takes the correlation structure among the forecasts $\hat{y}_{t+h}^{(i)}$ into the calculation of the weights, as it is based on the regression in Equation (7.6), just as in the regression in Equation (7.3) to get Equation (7.2).

FURTHER TOPICS IN COMBINING FORECASTS

We have discussed the combining forecasts for one-step-ahead forecasting, for the conditional mean, of continuous random variables. This can be extended to the following three things:

1) multiple-step-ahead forecasts;
2) conditional variance forecasts, conditional quantile forecasts, conditional density forecasts, and conditional interval forecasts; and
3) discrete random variables (categorized data, binary data).

Combining Multistep Forecasting

Lin and Granger (1994) classify the multistep mean forecast methods into five alternative categories. Let's assume the true DGP can be characterized by the following equation:

$$Y_{t+1} = g\left(Y_t\right) + \varepsilon_{t+1} ,$$

where ε_t is a zero-mean, independent, and identically distributed sequence with distribution function Φ.

The optimal one-step forecast using a least square criterion is

$$Y_{t,1} = E\left[Y_{t+1} \middle| Y_{t-j},\, j \geq 0 \right] = g\left(Y_{t-1}\right) .$$

When $g(\cdot)$ is known, there should be no problem in generating a one-step-ahead forecast. When $g(\cdot)$ is not known in practice, we can approximate $g(\cdot)$ by a flexible function form such as the polynomial family or the neural network family. However, the multistep forecasts for nonlinear models are much more complicated than the one-step forecast. Consider the simplest $h = 2$ case as an example to illustrate the multistep forecast methods. The optimal two-step-ahead forecast at time t is as follows:

$$\begin{aligned} Y_{t,2} &= E\left[Y_{t+2} \middle| Y_{t-j}, \quad j \geq 0 \right] \\ &= E\left[g\left(Y_{t+1}\right) + \varepsilon_{t+2} \middle| Y_{t-j}, \quad j \geq 0 \right] \\ &= E\left[g\left(g\left(Y_t\right) + \varepsilon_{t+1}\right) \middle| Y_{t-j}, \quad j \geq 0 \right] . \end{aligned} \tag{7.7}$$

Of the five multistep mean forecast methods, there are four possible ways to do multistep forecasts by iterating one-step-ahead forecasts, as is discussed by Brown and Mariano (1989):

1) Naive (or deterministic):

$$Y_{t,2}^{n} \equiv g\left(g\left(Y_t\right)\right),$$

so that the presence of ε_{t+1} is ignored by putting its value at zero. For most nonlinear function $g(\cdot)$, $Y_{t,2}^{n}$ will be biased, and the direction of the bias depends on whether $g(\cdot)$ is convex or concave, as discussed by Granger and Newbold (1976).

2) Exact (or optimal, or closed form):

$$Y_{t,2}^{e} \equiv \int_{-\infty}^{\infty} g\left(g\left(Y_t\right)+\varepsilon_{t+1}\right) d\boldsymbol{\Phi} .$$

3) Monte Carlo:

$$Y_{t,2}^{m} \equiv \frac{1}{J}\sum_{j=1}^{J} g\left(g\left(Y_t\right)+\varepsilon_j\right),$$

where $\varepsilon_j = 1, \ldots, J$ are random numbers drawn from the distribution Φ. If J is large enough, $Y_{t,h}^{m}$ and $Y_{t,h}^{e}$ should be virtually identical.

4) Bootstrap (or residual-based):

$$Y_{t,2}^{b} \equiv \frac{1}{t}\sum_{j=1}^{t} g\left(g\left(Y_t\right)+\hat{\varepsilon}_j\right),$$

where $\hat{\varepsilon}_j$, $j = 1, \ldots, t$ are the t values of the residual estimated over the sample period.

An alternative way of doing a multistep mean forecast is to model the relationship between Y_{t+h} and Y_t *directly* by a new function $g_h(\cdot)$:

5) $Y_{t+h} = g_h\left(Y_t\right) + e_{t,h}$,

though $e_{t,h}$ is usually not white noise, as mentioned by Lin and Granger (1994). Therefore, a fifth method for doing a multistep forecast is

$$Y_{t,h}^{d} \equiv g_h\left(Y_t\right) .$$

With any of these five methods, the factor models considered in the previous sections may be used for multistep forecasts when there are many predictors or many forecasts.

Combining Quantile Forecasts

The optimal forecast $\hat{y}_{t+1}$ may be estimated, for a given $\alpha \in (0,1)$, from minimizing the check loss:

$$\min_{\hat{y}_{t+1}} \rho_\alpha\left(e_{t+1}\right) = \left[\alpha - 1\left(e_{t+1} < 0\right)\right] \times e_{t+1},$$

where $e_{t+1} = y_{t+1} - \hat{y}_{t+1}$. Since $\rho_\alpha(\cdot)$ is convex, the results of Bates and Granger (1969), as discussed in the next section, can be carried over.

Note that the optimal forecast $\hat{y}^*_{t+1} = q_\alpha\left(y_{t+1} \middle| \boldsymbol{x}_t\right)$ satisfies the following first-order condition:

$$E\left(\alpha - \mathbf{1}\left(y_{t+1} < \hat{y}^*_{t+1}\right) \middle| \boldsymbol{x}_t\right) = 0, \qquad \text{a.s.}$$

(See, e.g., Giacomini and Komunjer [2005].) Hence,

$$g_{t+1} \equiv \alpha - \mathbf{1}\left(y_{t+1} < \hat{y}^*_{t+1}\right)$$

may be called the generalized residual or generalized forecast error. From this we obtain

$$\alpha = E\left(\mathbf{1}\left(y_{t+1} < \hat{y}^*_{t+1}\right) \middle| \boldsymbol{x}_t\right) = Pr\left(y_{t+1} \leqq \hat{y}^*_{t+1} \middle| \boldsymbol{x}_t\right).$$

It is interesting to note that this corresponds exactly to Equation (7.8) on page 169 for evaluating interval forecasts, whereas here we apply it to the optimal forecast $\hat{y}^*_{t+1}$.

We consider two types of data-rich environments—one where there are *N* predictors,

$$\boldsymbol{x}_t = \left(x_{t1}, x_{t2}, \ldots, x_{tN}\right)',$$

and another where there are *N* quantile forecasts, with

$$\hat{\boldsymbol{y}}_{t+h} = \left(\hat{y}^{(1)}_{t+h}, \hat{y}^{(2)}_{t+h}, \ldots, \hat{y}^{(N)}_{t+h}\right)'.$$

It is necessary to generalize the principal component regression for conditional quantiles under the check loss $\rho_\alpha(\cdot)$.

Combining Density Forecasts

Suppose that $\{y_t\}_{t=-\infty}^{\infty}$ is a time series (e.g., the return of a portfolio over a certain period) with unknown conditional density function $f_t(y) \equiv f_t(y|\boldsymbol{x}_{t-1})$. Let $p_t(y,\theta) \equiv p_t(y|\boldsymbol{x}_{t-1},\theta)$ be a one-step-ahead conditional density forecast model, where θ is a finite-dimensional parameter. Suppose that $p_t(y,\theta_0) = f_t(y)$ for some θ_0. Then, show that the one-step-ahead density forecast is optimal in the sense that it dominates all other density forecasts for any loss function (Diebold, Gunther, and Tay 1998; Granger 1999; Granger and Pesaran 2000). In practice it is not uncommon that a suboptimal forecast model does better than another in predicting a certain aspect of the distribution (e.g., value at risk at the 5 percent level) but worse than another in predicting a different aspect of the distribution (e.g., value at risk at the 1 percent level). This makes it difficult for forecast users (who may not be forecast producers) to choose a suitable forecast model. The fact that the optimal forecast model is preferred by all forecast users regardless of their loss functions resolves this difficulty. It is therefore useful to check whether a density forecast model is optimal, and, if not, to determine what useful information can be provided from it for further improvement in density forecasts. In fact, even if point forecasts are of interest, the optimal conditional density forecasts are needed to construct optimal point forecasts under a general asymmetric loss function (Christoffersen and Diebold 1996, 1997).

Suppose that $\{y_t\}$ is generated from conditional densities $\{f_t(y)\}$. If a sequence of density forecasts $\{p_t(y,\theta_0)\}$ coincides with $\{f_t(y)\}$, then under the usual condition of a nonzero Jacobian with continuous partial derivatives, $\{Z_t\}$ is IID $U[0,1]$. That is, when the forecast model $p_t(y,\theta)$ is optimal, the series of PITs, $\{Z_t\}$, where

$$Z_t \equiv \int_{-\infty}^{y_t} p_t(y,\theta_0)\,dy,$$

is IID $U[0,1]$. (See Diebold, Gunther and Tay [1998].) Berkowitz (2001) considers the inverse normal transformation of the PIT, which follows IID $N(0,1)$. Bao, Lee, and Saltoglu (2007) discuss how the Kullback-

Leibler Information Criterion (KLIC) of Kullback and Leibler (1951), based on PIT, may be used to compare the density forecasts. (See Mitchell and Hall [2005] for combining density forecasts.) Combining many density forecasts (with large N),

$$\left(\hat{f}_{t+h}^{(1)}(y), \hat{f}_{t+h}^{(2)}(y), \ldots, \hat{f}_{t+h}^{(N)}(y)\right)',$$

would require combinations of conditional moments or conditional quantiles, with mixtures of several distributions, which would be complicated.

Combining Interval Forecasts

Consider a stationary series $\{y_t\}_{t=1}^{T}$. Let the one-period-ahead conditional interval forecast made at time t from a model be denoted as

$$J_{t,1}(\alpha) = \left(L_{t,1}(\alpha), U_{t,1}(\alpha)\right), \qquad t = R, \ldots, T,$$

where $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ are the lower and upper limits of the ex ante interval forecast for time $t+1$ made at time t with the coverage probability α, i.e., $\alpha = Pr\left[y_{t+1} \in J_{t,1}(\alpha) | \boldsymbol{x}_t\right]$. If we define the indicator variable as $d_{t+1}(\alpha) = \mathbf{1}\left[y_{t+1} \in J_{t,1}(\alpha)\right]$, the sequence $\{d_{t+1}(\alpha)\}_{t=R}^{T}$ is IID Bernoulli (α). The optimal interval forecast would satisfy

$$(7.8) \quad E\left(d_{t+1}(\alpha) | \boldsymbol{x}_t\right) = \alpha,$$

so that $\{d_{t+1}(\alpha) - \alpha\}$ will be a martingale difference sequence. As the $\{d_{t+1}(\alpha)\}$ has the expected Bernoulli log-likelihood

$$E\alpha^{d_{t+1}(\alpha)}(1-\alpha)^{\left[1-d_{t+1}(\alpha)\right]},$$

we can choose a model with the largest out-of-sample mean of

$$p^{-1}\sum_{t=R}^{T}\log\left(\alpha^{\hat{d}_{t+1}(\alpha)}\left[1-\alpha\right]^{\left[1-\hat{d}_{t+1}(\alpha)\right]}\right).$$

(See Bao, Lee, and Saltoglu [2006].)

To combine interval forecasts that are generated from multiple models, one can use the conditional quantile forecasts derived from using regression quantiles for $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ and combine them; or one can use the conditional density forecasts, combine them, and invert the combined density forecast to get the conditional quantile forecasts for $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$, using the methods discussed in Section 9.2.

Combining Binary Forecasts

Lee and Yang (2006) consider binary forecasts using bagging to form a (weighted) average over all bootstrap training samples drawn from the same distribution. The idea can be extended to cases where there are many predictors or many forecasts to form a combined forecast of many binary forecasts. As in Lee and Yang, the combined binary predictor $\hat{y}_t^{(c)}$ can be constructed by the majority voting on the N individual binary forecasts $\hat{y}_t^{(i)}$ ($i = 1, \ldots, N$), i.e.,

$$\hat{y}_t^{(c)} = \mathbf{1}\left(\sum_{i=1}^{N} w_i \hat{y}_t^{(i)} > \frac{1}{2} \right),$$

where $\sum_{i=1}^{N} w_i = 1$. It is not clear how to estimate the combination weights $\{w_i\}$ when N is large. Simple cases are those where we assume a perfect democracy, with $w_i = 1/N$ for all i, and those where we assume a dictator, with $w_i = 1$ for some i. Neither case can be optimal in terms of the binary loss functions.

CONCLUSIONS

We have considered how to combine forecasts in a data-rich environment with many predictors, $\boldsymbol{x}_t = \left(x_{t1}, x_{t2}, \ldots, x_{tN}\right)'$, or with many forecasts, $\hat{\boldsymbol{y}}_{t+h} = \left(\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \ldots, \hat{y}_{t+h}^{(N)}\right)'$ (when N is large). In practice there are situations where we, whether econometricians or forecasters, do not observe the predictors but only the forecasts (e.g., survey forecasts of the Federal Reserve Bank of Philadelphia). In such situations one needs to aggregate many forecasts into a consensus group forecast. A common way is to use the simple average, or majority voting. While

many empirical results from out-of-sample forecasting have shown that the simple average of multiple forecasts tends to work well, such a conclusion assumes that all individual forecasts are equally good by assigning equal weights. The accuracy can be improved if the weights can be estimated consistently without experiencing errors from the usual large *N* problem (the so-called curse of dimensionality). We use a factor model of many forecasts to derive the forecast combination weights without succumbing to this problem.

In a data-rich environment with many predictors or many forecasts, it is often necessary to use reduced-dimension specifications that can span a large number of predictors. In the recent forecasting literature, the use of factor models and principal component estimation has been advocated for forecasting in the presence of many predictors. In this situation, we decompose the space spanned by many predictors using the principal components, as in Stock and Watson (2002). We can also project the forecast target to many subspaces spanned by the predictors, obtain many artificially generated forecasts, and then combine those forecasts generated from the subspaces, as in Chan, Stock, and Watson (1999); Hillebrand et al. (2010); and Tu and Lee (2009).

References

Aiolfi, Marco, and Alan Timmermann. 2006. "Persistence in Forecasting Performance and Conditional Combination Strategies." *Journal of Econometrics* 135(1–2): 31–53.

Ang, Andrew, and Monika Piazzesi. 2003. "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables." *Journal of Monetary Economics* 50(4): 745–787.

Ang, Andrew, Monika Piazzesi, and Min Wei. 2006. "What Does the Yield Curve Tell Us about GDP Growth?" *Journal of Econometrics* 131(1–2): 359–403.

Bai, Jushan. 2003. "Inferential Theory for Factor Models of Large Dimensions." *Econometrica* 71(1): 135–171.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70(1): 191–221.

———. 2006. "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions." *Econometrica* 74(4): 1133–1150.

Bao, Yong, Tae-Hwy Lee, and Burak Saltoglu. 2006. "Evaluating Predictive

Performance of Value-at-Risk Models in Emerging Markets: A Reality Check." *Journal of Forecasting* 25(2): 101–128.

———. 2007. "Comparing Density Forecast Models." *Journal of Forecasting* 26(3): 203–225.

Bates, J.M., and Clive W.J. Granger. 1969. "The Combination of Forecasts." *Operational Research Quarterly* 20(4): 451–468.

Berkowitz, Jeremy. 2001. "Testing Density Forecasts, with Applications to Risk Management." *Journal of Business and Economic Statistics* 19(4): 465–474.

Bernanke, Ben S. 1990. "On the Predictive Power of Interest Rates and Interest Rate Spreads." Federal Reserve Bank of Boston's *New England Economic Review* 1990(November/December): 51–68.

Bernanke, Ben S., and Jean Boivin. 2003. "Monetary Policy in a Data-Rich Environment." *Journal of Monetary Economics* 50(3): 525–546.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(2): 123–140.

Brown, Bryan W., and Roberto S. Mariano. 1989. "Predictors in Dynamic Nonlinear Models: Large-Sample Behavior." *Econometric Theory* 5(3): 430–452.

Chan, Yeung Lewis, James H. Stock, and Mark W. Watson. 1999. "A Dynamic Factor Model Framework for Forecast Combination." *Spanish Economic Review* 1(2): 91–121.

Christoffersen, Peter F., and Francis X. Diebold. 1996. "Further Results on Forecasting and Model Selection under Asymmetric Loss." *Journal of Applied Econometrics* 11(5): 561–572.

———. 1997. "Optimal Prediction under Asymmetric Loss." *Econometric Theory* 13(6): 808–817.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5(4): 559–583.

Deutsch, Melinda, Clive W.J. Granger, and Timo Teräsvirta. 1994. "The Combination of Forecasts Using Changing Weights." *International Journal of Forecasting* 10(1): 47–57.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39(4): 863–883.

Diebold, Francis X., and Peter Pauly. 1990. "The Use of Prior Information in Forecast Combination." *International Journal of Forecasting* 6(4): 503–508.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." With discussion. *Annals of Statistics* 32(2): 407–499.

Gatev, Evan, William N. Goetzmann, and K.Geert Rouwenhorst. 2006. "Pairs

Trading: Performance of a Relative-Value Arbitrage Rule." *Review of Financial Studies* 19(3): 797–827.

Giacomini, Raffaella, and Ivana Komunjer. 2005. "Evaluation and Combination of Conditional Quantile Forecasts." *Journal of Business and Economic Statistics* 23(4): 416–431.

Granger, Clive W.J. 1999. *Empirical Modeling in Economics: Specification and Evaluation*. Cambridge: Cambridge University Press.

Granger, Clive W.J., and Yongil Jeon. 2004. "Thick Modelling." *Economic Modelling* 21(2): 323–343.

Granger, Clive W.J., and Paul Newbold. 1976. "Forecasting Transformed Series." *Journal of the Royal Statistical Society*, Series B, 38(2): 189–203.

———. 1986. *Forecasting Economic Time Series,* 2d ed. New York: Academic Press.

Granger, Clive W.J., and M. Hashem Pesaran. 2000. "Economic and Statistical Measures of Forecast Accuracy." *Journal of Forecasting* 19(7): 537–560.

Granger, Clive W.J., and Ramu Ramanathan. 1984. "Improved Methods of Combining Forecasts." *Journal of Forecasting* 3(2): 197–204.

Hansen, Bruce E. 2008. "Least-Squares Forecast Averaging." *Journal of Econometrics* 146(2): 342–350.

Hillebrand, Eric, Tae-Hwy Lee, Canlin Li, and Huiyu Huang. 2010. "Forecasting Output Growth and Inflation: How to Use Information in the Yield Curve." Working paper. Riverside, CA: University of California, Riverside.

Hong, Yongmiao, and Tae-Hwy Lee. 2003. "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models." *Review of Economics and Statistics* 85(4): 1048–1062.

Huang, Huiyu, and Tae-Hwy Lee. 2010. "To Combine Forecasts or to Combine Information?" *Econometric Reviews* 29(5–6): 534–570.

Kullback, S., and R.A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22(1): 79–86.

Lee, Tae-Hwy, Canlin Li, and Huiyu Huang. 2010. "Pairs Trading Strategy for Combining Forecasts." Working paper. Riverside, CA: University of California, Riverside.

Lee, Tae-Hwy, and Yang Yang. 2006. "Bagging Binary and Quantile Predictors for Time Series." *Journal of Econometrics* 135(1–2): 465–497.

Lin, Jin-Lung, and Clive W.J. Granger. 1994. "Forecasting from Non-Linear Models in Practice." *Journal of Forecasting* 13(1): 1–9.

Mitchell, James, and Stephen G. Hall. 2005. "Evaluating, Comparing, and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation." *Oxford Bulletin of Economics and Statistics* 67(S1): 995–1033.

Newbold, Paul, and Clive W.J. Granger. 1974. "Experience with Forecasting Univariate Time Series and the Combination of Forecasts." *Journal of the Royal Statistical Society* 137(2): 131–165.

Palm, Franz C., and Arnold Zellner. 1992. "To Combine or Not to Combine? Issues of Combining Forecasts." *Journal of Forecasting* 11(8): 687–701.

Shen, Xiaotong, and Hsin-Cheng Huang. 2006. "Optimal Model Assessment, Selection, and Combination." *Journal of the American Statistical Association* 101(474): 554–568.

Smith, Jeremy, and Kenneth F. Wallis. 2009. "A Simple Explanation of the Forecast Combination Puzzle." *Oxford Bulletin of Economics and Statistics* 71(3): 331–355.

Stock, James H., and Mark W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators." In *NBER Macroeconomics Annual 1989*, Vol. 4, Olivier Jean Blanchard and Stanley Fischer, eds. Cambridge, MA: MIT Press, pp. 351–408.

———. 1996. "Evidence on Structural Instability in Macroeconomic Time Series Relations." *Journal of Business and Economic Statistics* 14(1): 11–30.

———. 1999. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, Robert F. Engle and Halbert White, eds. Oxford: Oxford University Press, pp. 1–45.

———. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97(460): 1167–1179.

———. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set." *Journal of Forecasting* 23(6): 405–430.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society*, Series B, 58(1): 267–288.

Timmermann, Allan. 2006. "Forecast Combinations." In *Handbook of Economic Forecasting*, Vol. 1, Graham Elliott, Clive W.J. Granger, and Allan Timmermann, eds. Amsterdam: North-Holland, pp. 135–196.

Tu, Yundong, and Tae-Hwy Lee. 2009. "Forecasting Using Supervised Factor Models." Working paper. Riverside, CA: University of California, Riverside.

Wright, Jonathan H. 2009. "Forecasting U.S. Inflation by Bayesian Model Averaging." *Journal of Forecasting* 28(2): 131–144.

Yang, Yuhong. 2004. "Combining Forecasting Procedures: Some Theoretical Results." *Econometric Theory* 20(1): 176–222.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society*, Series B, 67(2): 301–320.