Combining Forecasts with Many Predictors

Tae-Hwy Lee*
Department of Economics
University of California, Riverside

August 16, 2010

Abstract

We discuss how we form a combined forecast in data rich environment where there are many predictors or many forecasts. It is often necessary to use reduced-dimension specifications that can span a large number of predictors or forecasts. The use of factor models and principal component estimation has been advocated for forecasting in the presence of many predictors. We decompose the space spanned by many predictors using principal components. Or we can project the forecast target to many subspaces spanned by the predictors, then obtain many artificially generated forecasts, and then combine those forecasts generated from the subspaces. Finally, we discuss some issues in forecast combination for quantile forecasts, density forecasts, interval forecasts, and binary forecasts.

Key Words: Combining forecasts, Data rich environment, Many predictors, Many forecasts, Forecast combination weights, Principal components.
JEL Classification: C5, E4, G1.

*Department of Economics, University of California, Riverside, CA 92521-0427, U.S.A. Phone: +1 (951) 827-1509. E-mail: taelee@ucr.edu.

1 Introduction

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). 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 (BMA) may be used to form a weighted combined forecast. See e.g., Lee and Yang (2006). A frequently asked question is: "how to combine". See, for example, 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* modelling and write, "An advantage of thick modelling 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 usually in practice 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 over the original forecast of the model. This is the idea of bootstrap aggregating (or bagging), introduced by Breiman (1996).

Huang and Lee (2010) consider the situation when one wants to predict an economic

variable using the information set of many relevant explanatory variables. Diebold and Pauly (1990) 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, e.g., 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 the case of professional forecasters) while the underlying information set and the model used for generating each individual forecast are unknown.

In this paper we consider how to combine forecasts in a situation where *many* predictors (i.e., 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), Stock and Watson (2002), Huang and Lee (2010), Hillebrand, Lee, Li, and Huang (2010), and Tu and Lee (2010)

2 Data rich environment

Bernanke and Boivin (2003) emphasize that the use of large data set is common environment in practice such as in the central bank's policy making analysis. They wrote, "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."

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

Bernanke and Boivin (2003) confirm merit of the large data set as follows: "... 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 reconfirm Stock and Watson's result that the use of large data sets can improve forecast accuracy ..."

A natural question is how we should use all those vast data in predicting a target of interest. Using large data, there are advantages of accessing rich information and robustifying against structural instability which plagues low dimensional forecasting. While we can exploit these advantages, there are also difficulties of 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, known as a factor model, has recently been a popular approach in econometric forecasting, due to 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, known as variable selection, has been widely studied in statistics.

The variable selection is to reduce N by ranking and selecting a subset of X that are most predictive for a forecast target y, via such methods as LASSO (Tibshirani 1996), least angle regression (Efron, Hastie, Johnstone and Tibshirani 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 analysts in an investment bank, many different government agents, etc. In this paper we consider both cases, namely, the data rich environment with many predictors with N-vector $\mathbf{x}_t = (x_{t1},\ x_{t2},\ \ldots,\ x_{tN})'$, and the data rich environment with many forecasts with N-vector of forecasts $\hat{\mathbf{y}}_{t+h} = (\hat{y}^{(1)}_{t+h},\ \hat{y}^{(2)}_{t+h},\ \ldots,\ \hat{y}^{(N)}_{t+h})'$. Below we will discuss how we form a forecast under these two types of data rich environment. In both cases, the idea is to combining 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 forecasts even if it's a dominate 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 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 with $N = 2$.

3 Combining forecasts

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

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

such that $Ee_{t+1}^{(i)} = 0$, $Ee_{t+1}^{(i)2} = \sigma_i^2$, 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

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

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) showed that the optimal combined forecast error loss is smaller than the smaller of the two individual forecast error losses

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

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, committees of learners, a team of forecasts, consensus of learners, mixture of experts, expert system, and etc.

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

Why do we combine? Aiolfi and Timmermann (2006) have given answers as follows, "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."

Aiolfi and Timmermann (2006) also wrote:

1. Forecasts are of considerable importance to decision makers throughout economics and finance and are routinely used by private enterprises, government institutions and professional economists. 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 and even the most complicated model is likely to be misspecified and can, at best, provide a reasonable 'local' approximation to the target variable. Particularly so in practical forecasting situations in macroeconomics with a large cross-section of forecasting models and a short time-series dimension.

2. Forecasting models are also likely to be subject to structural instability as documented by Stock and Watson (1996) for a large series of macroeconomic variables. Viewed this way, it is highly unlikely that a single model will dominate all others at all points in time and the identity of the best 'local' approximation is also likely to change over time. If the identity of the best 'local' model is time-varying, a question immediately arises, namely whether it is possible to design a forecasting strategy that, at each point in time, selects the best current model. This strategy can be expected to be successful if the best current forecasting model outperforms all other models by a suitably large margin and if the ranking of the forecasting models is persistent. If either of these conditions fails to be met, a strategy of selecting only a single 'best' model is unlikely to work well. Most obviously, if (ex-ante) the identity of the best model varies randomly from period

to period, it will not be possible to identify this model by considering past forecasting performance across models. Similarly, if the best model 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 likely that diversification gains from combining across a set of forecasting models with similar performance will dominate the strategy of only using a single model.

5 How to combine?

The optimal combination weights in (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 (2) may suffer from the 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 using portfolio variance minimization methods. 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 because 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) used ranking of the forecasting model and 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. Also popular method is Bayesian model averaging (BMA) used in many applications, e.g., Lee and Yang (2006) and Wright (2009).

The formula for the optimal combination weights in (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 needs 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 (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 over one 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 outperform any individual forecast and explains why an inferior forecast can be included in the combination to improve forecast. The optimal combination weight has a pattern 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 multi-step 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

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

which can be obtained from the regression

$$e_{t+1}^{(2)} = w \left(e_{t+1}^{(2)} - e_{t+1}^{(1)} \right) + e_{t+1}^{(c)}. \tag{3}$$

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 one) 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 ignoring 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 the "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 lets assume $\hat{y}_t^{(1)}$ 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 $\hat{w}_t'$ ignoring the correlation ρ would be too restrictive.

6 Forecasting in data rich environment

So far, we have considered the case when $N = 2$. Most of the combining forecast literature has been limited to the case when N is small. In the present paper we consider the combining forecasts when N is large. Consider a kitchen-sink model with all predictors $\mathbf{x}_t$ in one large model

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

to generate the h-step forecast

$$\hat{y}_{T+h} = (1 \ \mathbf{x}_T')\hat{b}_T.$$

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 $MSFE = \mathbb{E}(y_{t+h} - \hat{y}_{t+h})^2 = O\left(\frac{N}{T}\right)$. 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 factor model to reduce N, or both. The variable selection is to reduce N by ranking variables in X and selecting a subset of X that are most predictive for a forecast target y, via such methods as forward and backward selection, stepwise regression, LASSO (Tibshirani 1996), least angle regression (Efron 2004), elastic net (Zou and Hastie 2005), and so on.

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

$$\begin{aligned} y_{t+1}^{(1)} &= x_{1t}\beta_1 + \varepsilon_{1,t+1} \\ &\vdots \\ y_{t+1}^{(N)} &= x_{Nt}\beta_N + \varepsilon_{N,t+1}, \end{aligned}$$

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

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

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

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

and the second is where there are N forecasts with

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

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

7 Forecasting with many predictors

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

$$\mathbf{x}_t = \Lambda f_t + v_t, \tag{4}$$

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

$$y_t = f_t\alpha + u_t. \tag{5}$$

As noted in Hillebrand, Lee, Li, and Huang (2010) and Tu and Lee (2010), in this approach, the factors are obtained from the marginal model of $\mathbf{x}_t$ rather than the joint model of $(y\ \ \mathbf{x}_t)$. Write the above model in (4) and (5) as follows

$$\begin{aligned} y_t &= \mathbb{E}(y_t|\mathbf{x}_t;\theta_1) + u_t = f_t\alpha + u_t \qquad (\theta_1 = \alpha) \\ \mathbf{x}_t &= \mathbb{E}(\mathbf{x}_t;\theta_2) + v_t = \Lambda f_t + v_t \qquad (\theta_2 = f_t, \Lambda). \end{aligned}$$

Note that this *assumes* that the joint density $D(y_t, \mathbf{x}_t; \theta) = D_1(y_t|\mathbf{x}_t; \theta) \cdot D_2(\mathbf{x}_t; \theta)$ operates a cut

$$D(y_t, \mathbf{x}_t; \theta) = D_1(y_t|\mathbf{x}_t; \theta_1) \cdot D_2(\mathbf{x}_t; \theta_2),$$

where $(\theta_1 \ \theta_2) \in \Theta_1 \times \Theta_2$ are variation free, and we estimate the conditional model (5) and the marginal model (4) separately.

8 Forecasting with many forecasts

Next, we consider forecasting when there are N forecasts $\hat{\mathbf{y}}_{t+h} = (\hat{y}^{(1)}_{t+h},\ \hat{y}^{(2)}_{t+h},\ \ldots,\ \hat{y}^{(N)}_{t+h})'$ and N is large. In this situation many forecasts are given either from many survey forecasters or from many analysts. There are various organizations which operate as an aggregate or a group, based on many individual analysts which 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

$$\begin{aligned} \hat{\mathbf{y}}_{t+h} &= \Lambda f_{t+h} + v_{t+h} \\ \hat{f}_{t+h} &= \hat{\Lambda}' \hat{\mathbf{y}}_{t+h} \end{aligned}$$

and estimate the following forecasting equation

$$y_{t+h} = \hat{f}'_{t+h}\alpha + u_{t+h} \tag{6}$$

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

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

and therefore the optimal forecast combination weights are

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

Hillebrand, Lee, Li, and Huang (2010) and Tu and Lee (2010) 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 $\hat{y}^{(i)}_{t+h}$ is generated. In practice, there are various situations where only 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 US 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 often 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 are when N is not very large. When N is very large, the equal weights $\left(\frac{1}{N}\right)$ put too little weights to good models, especially when $\frac{1}{N} \to \infty$, and the equal weights can be hardly 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 $\hat{y}^{(i)}_{t+h}$ into the calculation of the weights as it is based on the regression in (6) just like in the regression (3) to get (2).

9 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:

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

9.1 Combining multi-step forecasting

Lin and Granger (1994) summarize the multi-step mean forecast methods into five alternative categories. Assume the true DGP can be characterized by the following equation:

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

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:

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

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

ahead forecast at time t is:

$$\begin{aligned} Y_{t,2} &= \mathbf{E}[Y_{t+2}|Y_{t-j},\ j \geq 0] \\ &= \mathbf{E}[g(Y_{t+1}) + \varepsilon_{t+2}|Y_{t-j},\ j \geq 0] \\ &= \mathbf{E}[g(g(Y_t) + \varepsilon_{t+1})|Y_{t-j},\ j \geq 0]. \end{aligned} \tag{7}$$

There are four possible ways to do multi-step forecasts by iterating one-step ahead forecasts as also discussed by Brown and Mariano (1989):

1. **naive (or deterministic)**

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

so that the presence of ε_{t+1} is ignored by putting its value to zero. For most non-linear 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)**

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

3. **Monte Carlo**

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

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

4. **bootstrap (or residual based)**

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

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

An alternative way for multi-step mean forecast is to model the relationship between Y_{t+h} and Y_t *directly* by a new function $g_h(\cdot)$

$$Y_{t+h} = g_h(Y_t) + e_{t,h},$$

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

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

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

9.2 Combining quantile forecasts

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

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

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 Section 3 can be carried over. The proof of the results are left for exercise.

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

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

See e.g., Giacomini and Komunjer (2005). Hence, $g_{t+1} \equiv \alpha - \mathbf{1}(y_{t+1} < \hat{y}_{t+1}^*)$ may be called as the generalized residual or generalized forecast error. From this, we obtain

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

It is interesting to note that this corresponds exactly to the equation (8) in Section 9.4 for evaluating interval forecasts, while here we apply to the optimal forecast $\hat{y}_{t+1}^*$.

We can consider two types of data rich environments – one where there are N predictors

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

and another where there are N quantile forecasts with

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

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

9.3 Combining density forecast

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 $f_t(y) \equiv f_t(y|\mathbf{x}_{t-1})$. Let $p_t(y,\theta) \equiv p_t(y|\mathbf{x}_{t-1},\theta)$ be a 1-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 1-step-ahead density forecast is optimal in the sense that it dominates all other density forecasts for any loss function (Diebold *et al.* 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% level) but vice versa in predicting another aspect of the distribution (e.g., value at risk at the 1% level). This makes forecast users (who may not be forecast producers) difficult 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 important and useful to check whether a density forecast model is optimal, and if not, what useful information can be provided 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) considered the inverse normal transform of the PIT which follows IID $N(0,1)$. Bao, Lee and Saltoglu (2007) discuss how 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)

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

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

9.4 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

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

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., $\alpha = \Pr[y_{t+1} \in J_{t,1}(\alpha)|\mathbf{x}_t]$. If we define the indicator variable $d_{t+1}(\alpha) = \mathbf{1}[y_{t+1} \in J_{t,1}(\alpha)]$, the sequence $\{d_{t+1}(\alpha)\}_{t=R}^T$ is IID Bernoulli (α). The optimal interval forecast would satisfy

$$E(d_{t+1}(\alpha)|\mathbf{x}_t) = \alpha, \tag{8}$$

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

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

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

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

See Bao, Lee, and Saltoglu (2006).

To combine interval forecasts that are generated from multiple models, one can use the conditional quantile forecasts 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.

9.5 Combining binary forecasts

Lee and Yang (2006) consider a 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 the cases where there are many predictors or many forecasts to form a combined forecast of many binary forecasts. As in Lee and Yang (2006), the combined binary predictor $\hat{y}_t^{(c)}$ can be constructed by the majority voting over 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 where we can assume a perfect democracy with $w_i = \frac{1}{N}$ for all i, and where we assume a dictator with $w_i = 1$ for some i. Neither cases can be optimal in terms of the binary loss functions.

10 Conclusions

We have considered how to combine forecasts in data rich environment with many predictors $\mathbf{x}_t = (x_{t1},\ x_{t2},\ \ldots,\ x_{tN})'$ or with many forecasts $\hat{\mathbf{y}}_{t+h} = (\hat{y}^{(1)}_{t+h},\ \hat{y}^{(2)}_{t+h},\ \ldots,\ \hat{y}^{(N)}_{t+h})'$ (when N is large). In practice there are situations where we, econometricians or forecasters, do not observe the predictors but only the forecasts (e.g., survey forecasts of Philadelphia Federal Reserve Bank). 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 on out-of-sample forecasting have shown that the simple average of multiple forecasts tends to work well, it assumes that all individual forecasts are equally good by assigning the equal weights. It can be improved if the weights can be estimated consistently and without suffering from the usual large N problem (so called the curse of dimensionality). We use a factor model of many forecasts to derive the forecast combination weights without suffering from the curse of dimensionality.

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 that situation we decompose the space spanned by many predictors using 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, Lee, Li, and Huang (2010), and Tu and Lee (2010)

11 References

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

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

Ang, A., Piazzesi, M, and Wei, M. (2006), "What Does the Yield Curve Tell Us about GDP Growth?" *Journal of Econometrics* 131, 359-403.

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

Bai, J. and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica* 70(1), 191-221.

Bai, J. and Ng, S. (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica* 74(4), 1133-1150.

Bai, J. and Ng, S. (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics* 146, 304-317.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association* 101:473, 119–137.

Bao, Yong, Tae-Hwy Lee, Burak Saltoglu (2006), "Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check", *Journal of Forecasting*, 25(2): 101-128.

Bao, Yong, Tae-Hwy Lee, Burak Saltoglu (2007), "Comparing Density Forecast Models", *Journal of Forecasting*, 26(3), 203-225.

Bates, J.M. and Granger, C.W.J. (1969), "The Combination of Forecasts", *Operations Research Quarterly* 20, 451-468.

Berkowitz, J. (2001), "Testing density forecasts with applications to risk management". *Journal of Business and Economic Statistics* 19: 465-474.

Bernanke, B.S. (1990), "On the Predictive Power of Interest Rates and Interest Rate Spreads," Federal Reserve Bank of Boston, *New England Economic Review*, 51-68.

Bernanke, B.S. and Jean Boivin (2003) "Monetary policy in a data-rich environmen", *Journal of Monetary Economics* 50: 525–546.

Breiman, L. (1996), "Bagging Predictors", *Machine Learning,* 24, 123-140.

Chan, Y.L., Stock, J.H., and Watson, M.W. (1999), "A Dynamic Factor Model Framework for Forecast Combination," *Spanish Economic Review* 1, 91-121.

Chong, Y.Y. and Hendry, D.F. (1986), "Econometric Evaluation of Linear Macro-Economic Models", *Review of Economics Studies,* LIII, 671-690.

Christoffersen, Peter. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.

Christofferson, Peter and Francis Diebold (2003), "Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics", McGill University and University of Pennsylvania.

Clark, T.E. and McCracken, M.W. (2009), "Combining Forecasts from Nested Models", *Oxford Bulletin of Economics and Statistics* 71, 303-329.

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

Deutsch, M., Granger, C.W.J., and Teräsvirta, T. (1994), "The Combination of Forecasts Using Changing Weights", *International Journal of Forecasting* 10, 47-57.

de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251-261

de Jong, S. and Kiers, H.A.L. (1992), "Principal Covariate Regression: Part I. Theory," *Chemometrics and Intelligent Laboratory Systems*, 14, 155-164.

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

Diebold, F.X., T. A. Gunther, and A. S. Tay (1998), "Evaluating density forecasts with applications to financial risk management", *International Economic Review* 39: 863-883.

Diebold, F.X. and Lopez, J.A. (1996), "Forecast Evaluation and Combination", NBER Working Paper, No. 192.

Diebold, F.X. and Pauly, P. (1990), "The Use of Prior Information in Forecast Combination", *International Journal of Forecasting*, 6, 503-508.

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

Engle, R.F., Granger, C.W.J. and Kraft, D.F. (1984), "Combining Competing Forecasts of Inflation Using a Bivariate ARCH Model", *Journal of Economic Dynamics and Control* 8, 151-165.

Engle, R.F., D.F. Hendry, J-F. Richard (1983), "Exogeneity", *Econometrica*, 51(2): 277-304

Estrella, A. (2005), "Why Does the Yield Curve Predict Output and Inflation?" *The Economic Journal* 115, 722-744.

Estrella, A. and Hardouvelis, G.A. (1991), "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance* 46, 555-76.

Fama, E. and Bliss, R. (1987), "The Information in Long-maturity Forward Rates," *American Economic Review* 77, 680-692.

Friedman, B.M., and Kuttner, K.N. (1991), "Why Does the Paper-Bill Spread Predict Real Economic Activity?" *NBER Working Paper*, No. 3879.

Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G. (2006), "Pairs Trading: Performance of a Relative-Value Arbitrage Rule," *Review of Financial Studies,* 19(3), 797-827.

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

Granger, C.W.J. (1989), "Invited Review: Combining Forecasts - Twenty years Later", *Journal of Forecasting* 8, 167-173.

Granger, C.W.J. and Ramanathan, R. (1984), "Improved Methods of Combining Forecasts", *Journal of Forecasting* 3, 197-204.

Granger, C.W.J. and Y. Jeon (2004), "Thick Modeling", *Economic Modeling*, 21, 323-343.

Granger, C.W.J. and P. Newbold (1976), "Forecasting Transformed Variables", *Journal of the Royal Statistical Society Series B*, 38, 189-203.

Granger, C.W.J. and P. Newbold (1986), *Forecasting Economic Time Series*, 2ed., Academic Press.

Granger, C.W.J. and M.H. Pesaran (2000), "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting,* 19, 537-560.

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

Hendry, D.F. and Clements, M.P. (2004), "Pooling of Forecasts", *Econometrics Journal* 7, 1-31.Granger, C.W.J. (1999), *Empirical Modeling in Economics: Specification and Evaluation,* Cambridge University Press: London.

Hibon, M. and Evgeniou, T. (2005), "To Combine or not to Combine: Selecting among Forecasts and Their Combinations", *International Journal of Forecasting* 21, 15-24.

Hillebrand, Eric, Tae-Hwy Lee, Canlin, Li, and Huiyu Huang (2010), "Forecasting Output Growth and Inflation: How to Use Information in the Yield Curve", UC 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 (2006), "To Combine Forecasts or to Combine Information?" in press (2010), *Econometric Reviews.*

Kullback, L. and R. A. Leibler (1951), "On information and sufficiency". *Annals of Mathematical Statistics* 22: 79-86.

Lancaster, T. (2000), "The incidental parameter problem since 1948," *Journal of Econometrics* 95: 391-413

Lee, Taehwy, Canlin, Li, and Huiyu Huang (2010), "Pairs Trading Strategy for Combining Forecasts", UC Riverside.

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

Lin, Jin-Lung, and C. W. J. Granger (1994), "Forecasting from non-linear models in practice", *Journal of Forecasting,* 13(1): pages...

Mitchell, J. and S. 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: 995-1033

Newbold, P. and Granger, C.W.J. (1974), "Experience with Forecasting Univariate Time Series and the Combination of Forecasts", *Journal of the Royal Statistical Society* 137, 131-165.

Newbold, P. and Harvey, D.I. (2001), "Forecast Combination and Encompassing", in *A Companion to Economic Forecasting*, Clements, M.P. and Hendry, D.F. (ed.), Blackwell Publishers.

Palm, F.C. and Zellner, A. (1992), "To Combine or not to Combine? Issues of Combining Forecasts", *Journal of Forecasting* 11, 687-701.

Shen, X. and Huang, H.-C. (2006), "Optimal Model Assessment, Selection, and Combination", *Journal of the American Statistical Association* 101, 554-568.

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

Stock, J.H. and Watson, M.W. (1989), "New Indexes of Coincident and Leading Indicators," *NBER Macroeconomic Annual*, Vol. 4, Olivier Blanchard and Stanley Fischer (ed.). Cambridge: MIT Press.

Stock, J.H. and Watson, M.W. (1999), "Forecasting Inflation," *Journal of Monetary Economics* 44, 293-335.

Stock, J.H. and M.W. Watson (1999), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series", in *Cointegration, Causality, and Forecasting, A Festschrift in Honor of C.W.J. Granger*, (eds.) R.F. Engle and H. White, Oxford University Press: London, pp. 1-44.

Stock, J.H. and Watson, M.W. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association* 97, 1167-1179.

Stock, J.H. and Watson, M.W. (2004), "Combination Forecasts of Output Growth in a Seven-country Data Set," *Journal of Forecasting* 23, 405-430.

Stock, J.H. and Watson, M.W. (2006), "Forecasting with Many Predictors", Chapter 10 in *Handbook of Economic Forecasting*, Volume 1, Elliott, G., Granger, C.W.J., and Timmermann, A. (ed.), North Holland.

Stock, J.H. and Watson, M.W. (2007), "Has Inflation Become Harder to Forecast?" *Journal of Money, Credit, and Banking* 39, 3-34.

Stock, J.H. and Watson, M.W. (2009), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," Harvard University and Princeton University.

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

Timmermann, A. (2006), "Forecast Combinations," *Handbook of Economic Forecasting*, Volume 1, Chapter 4, Elliott, G., Granger, C.W.J., and Timmermann, A. (ed.), North Holland.

Tu, Yundong and Tae-Hwy Lee (2010), "Forecasting Using Supervised Factor Models". UC Riverside.

Wright, J.H. (2009), "Forecasting US Inflation by Bayesian Model Averaging", *Journal of Forecasting* 28: 131-144.

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

Zou, Hui and Trevor Hastie (2005), "Regularization and variable selection via the elastic net", *J. R. Statist. Soc. B*, 67, Part 2, pp. 301–320.

Zou, H., T. Hastie, and Tibshirani, R. (2006), "Sparse Principal Component Analysis", *Journal of Computational and Graphical Statistics* 15(2): 262-286.