Forecast Combinations

Allan Timmermann*

UCSD

November 1, 2004

Abstract

Forecast combinations have frequently been found in empirical studies to produce better forecasts than methods based on the 'best' individual forecasting model. Moreover, simple combinations that ignore correlations between forecast errors often dominate more refined combination schemes aimed at estimating the theoretically optimal combination weights. In this chapter we analyze theoretically the factors that determine the advantages from combining forecasts (for example, the degree of correlation between forecast errors and the relative size of the individual models' forecast error variances). Although the reasons for the success of simple combination schemes are poorly understood, we discuss several possibilities related to model misspecification, instability (non-stationarities) and estimation error in situations where the numbers of models is large relative to the available sample size. We discuss the role of combinations under asymmetric loss and consider combinations of point, interval and probability forecasts.

*This research was sponsored by NSF grant SES0111238. I am grateful to Graham Elliott and Clive Granger for many discussions on the topic. Barbara Rossi provided detailed comments and suggestions that greatly improved the paper. Comments from seminar particpants at the UCSD Rady School forecasting conference were also helpful.

1 Introduction

Economists often have access to multiple forecasts of the same variable representing a variety of data sources or modeling approaches such as subjective judgements from experts or quantitative forecasts from linear or non-linear econometric models or from models with constant or time-varying parameters. Differences across forecasts could also reflect differences in individual forecasters' information sets due, for example, to their use of private information. Faced with multiple forecasts of the same variable, the question of how best to exploit information in the individual forecasts immediately arises. In particular, should a single dominant forecasting model be identified or should a combination of the underlying forecasts be used to produce a pooled summary measure? From a theoretical perspective, unless a particular forecasting model that generates smaller forecast errors than its competitors−with forecast errors that are uncorrelated with those from other models−can be identified *ex ante* forecast combinations offer diversification gains that make it attractive to combine individual forecasts rather than relying on forecasts from a single model. Even if the best model could be identified at each point in time, combination may still be an attractive strategy due to diversification gains, although its success will depend on how well the combination weights can be determined.

Forecast combinations have been used successfully in empirical work in diverse areas such as forecasting Gross National Product, currency market volatility, inflation, money supply, stock prices, meteorological data, city populations, outcomes of football games, wilderness area use, check volume and political risks, c.f. Clemen (1989). Summarizing the simulation and empirical evidence in the literature on forecast combinations, Clemen (1989, page 559) writes "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy.... in many cases one can make dramatic performance improvements by simply averaging the forecasts." More recently, Makridakis and Hibon (2000) conducted the so-called M3-competition which involved forecasting 3003 time series and concluded (p. 458) "The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods.". Similarly, Stock and Watson (2001, 2003) undertook an extensive study across

numerous economic and financial variables using linear and nonlinear forecasting models and found that pooled forecasts generally outperform predictions from the single best model, thus confirming Clemen's conclusion. Their analysis has been extended to a large European data set by Marcellino (2004) with broadly the same conclusions.

The classical argument for combining forecasts is provided in the seminal paper by Bates and Granger (1969). Its premise is that the information set underlying the individual forecasts is often unobserved to the forecast user. Hence the option of specifying a 'super' model that nests each of the underlying forecasting models and pooling their information sets is not feasible. For example, suppose that we are interested in forecasting some variable, Y, and that we are given two predictions, $\hat{Y}_1$ and $\hat{Y}_2$ of its conditional mean. Let the first forecast be based on the variables X_1, X_2, i.e., $\hat{Y}_1 = g_1(X_1, X_2)$, while the second forecast is based on the variables X_3, X_4, i.e., $\hat{Y}_2 = g_2(X_3, X_4)$. Further suppose that all variables enter with non-zero weights in the forecasts. If $\{X_1, X_2, X_3, X_4\}$ were observable, it would be natural to construct a forecasting model based on all four variables, $\hat{Y}_3 = g_3(X_1, X_2, X_3, X_4)$. On the other hand, if only the forecasts, $\hat{Y}_1$ and $\hat{Y}_2$ are observed by the forecast user−while the underlying variables are unobserved−then the only option is to combine these forecasts, i.e. to elicit a model of the type $\hat{Y} = g_c(\hat{Y}_1, \hat{Y}_2)$. More generally, the forecast user's information set, $\mathcal{I}$, may comprise n individual forecasts, $\mathcal{I} = \{\hat{Y}_1, ..., \hat{Y}_n\}$, but $\mathcal{I}$ is typically not the union of the information sets underlying the individual forecasts, $\cup_{i=1}^{n} \mathcal{I}_i$, but is a much smaller subset. Of course, the higher the degree of overlap in the underlying experts' information, the less useful a combination of forecasts is likely to be, c.f. Clemen (1987).

It is difficult to fully appreciate the strength of the diversification or hedging argument underlying forecast combination. Suppose the aim is to minimize some loss function belonging to a family of convex loss functions, $\mathcal{L}$, and that some forecast, $\hat{Y}_1$, stochastically dominates another forecast, $\hat{Y}_2$, in the sense that expected losses for all loss functions in $\mathcal{L}$ are lower under $\hat{Y}_1$ than under $\hat{Y}_2$. While this means that it is never rational for a decision maker to choose $\hat{Y}_2$ over $\hat{Y}_1$ in isolation, it is easy to construct examples where some combination of $\hat{Y}_1$ and $\hat{Y}_2$ generates a smaller expected loss than that produced using $\hat{Y}_1$ alone.

A second reason for using forecast combinations referred to by, inter alia, Figlewski

and Urich (1983), Kang (1986), Diebold and Pauly (1987), Makridakis (1989), Sessions and Chatterjee (1989), Winkler (1989), Hendry and Clements (2002) and Aiolfi and Timmermann (2004) and also thought of by Bates and Granger (1969) is that individual forecasts may be very differently affected by non-stationarities such as structural breaks caused, for example, by institutional change or technological developments. Some models may adapt quickly and will only temporarily be affected by structural breaks, while others only adapt more slowly as their parameters take longer to get updated using new post-break data. The longer the time spent since the most recent break, the better one might expect stable, slowly adapting models to perform relative to fast adapting ones as their parameters are more precisely estimated. If the time spent since the most recent break is rather short, the faster adapting models can be expected to produce the best forecasting performance. Since it is typically difficult to detect structural breaks in 'real time', it is plausible that on average, i.e., across periods with varying degrees of stability, combinations of forecasts from models with different degrees of adaptability will outperform forecasts from individual models.

A third and related reason for forecast combinations is that individual forecasting models will be subject to misspecification bias of unknown form, a point stressed particularly by Clemen (1989), Makridakis (1989), Diebold and Lopez (1996) and Stock and Watson (2001, 2003). Even in a stationary world, the data generating process is likely to be far more complicated than assumed by the most advanced model entertained by a forecaster. Viewing forecasting models as local approximations, it is unlikely that the same model dominates all others at all points in time. Rather, the best model will change over time in ways that can be difficult to track on the basis of past forecasting performance. Combining forecasts across different models can be viewed as a way to robustify the forecast against misspecification biases and measurement errors in the data sets underlying the individual forecasts. Notice the similarity to the classical portfolio diversification argument: Here the "portfolio" is the combination of forecasts, uncertainty is reflected in the forecast error and the source of risk reflects incomplete information about the target variable and model misspecification possibly due to non-stationarities in the underlying data generating process.

A fourth argument for combination of forecasts is that the underlying forecasts may be

based on different loss functions, c.f. Zellner (1986). This argument holds even if the forecasters observe the same information set. Suppose, for example, that forecaster A strongly dislikes large negative forecast errors while forecaster B strongly dislikes large positive forecast errors. In this case, forecaster A will under-predict the variable of interest (so the forecast error distribution is centered on a positive value), while forecaster B will over-predict it. If the bias is constant over time, there is no need to average across different forecasts since including a constant in the combination equation will pick up any unwanted bias. Suppose, however, that the optimal amount of bias is proportional to the conditional variance of the variable, as in Christoffersen and Diebold (1997). Provided that the two forecasters adopt a similar volatility model (which is not implausible since they are assumed to access to the same information set), a forecast user with a more symmetric loss function than was used to construct the underlying forecasts could find a combination of the two forecasts better than the individual ones.

Arguments against using forecast combinations are similarly numerous. Estimation errors that contaminate the combination weights are known to be a serious problem for many combination techniques, c.f. Diebold and Pauly (1990), Elliott (2004) and Yang (2004). Whereas non-stationarities in the underlying data generating process can be an argument for using combinations it can also lead to instabilities in the combination weights and cause great difficulty in deriving a set of combination weights that performs well, c.f. Clemen and Winkler (1986), Diebold and Pauly (1987), Figlewski and Urich (1983), Kang (1986) and Palm and Zellner (1992). In situations where the source of the different forecasts is a variety of individual forecasters whose (private) information sets are unobserved, most would agree that forecast combinations can add value. However, when the full set of predictor variables used to construct different forecasts is observed by the forecast user, it is more disputed whether a combination strategy should be used or whether a single best 'super' model that embeds all information should be searched for, c.f. Chong and Hendry (1986) and Diebold (1989).

If the advantages and disadvantages of forecast combinations seem familiar, this is not a coincidence. In fact, there are many similarities between the forecast combination problem

and the standard forecasting problem based on the construction of a single model. In both cases a subset of predictors (or individual forecasts) has to be selected among a larger set of potential forecasting variables and the choice of functional form mapping this information into the forecast as well as the choice of estimation method have to be determined. There are clearly important differences as well. First, when the information variables are themselves predictions of the target variable, it may be reasonable to assume that the individual predictors are unbiased in which case the combined forecast will also be unbiased provided that the combination weights are constrained to sum to unity and an intercept is omitted. If the unbiasedness assumption holds for each forecast, imposing such parameter constraints can lead to efficiency gains. Secondly, if the individual forecasts are generated by quantitative models whose parameters are estimated sequentially there is a potential generated regressor problem which could bias estimates of the combination weights. Third, since the individual forecasts are often based on information sets with considerable overlap, a multicollinearity problem arises in the sense that the individual weights are poorly estimated. In part this explains why using simple averages based on equal weights provides a natural benchmark. One would almost never want to impose equal weights on the coefficients of a standard regression model since explanatory variables can differ significantly in their units, interpretation and scaling. Finally, the forecasts that are being combined need not be point forecasts but could take the form of interval or density forecasts.

As a testimony to its important role in the forecasting literature, many high-quality surveys of forecast combinations have already appeared, c.f. Clemen (1989), Diebold and Lopez (1996) and Newbold and Harvey (2001). This survey differs from earlier ones in many important ways, however. First, we put more emphasis on the theory underlying forecast combinations, particularly in regard to the key diversification argument which is common also in portfolio analysis. Second, we deal in more depth with recent topics−some of which were emphasized as important areas of future research by Diebold and Lopez (1996)−such as combination of probability forecasts, combination under asymmetric loss and shrinkage.

The chapter is organized as follows. Given the importance of the underlying theory to interpret the vast amount of empirical work in the literature, we first cover the theoretical

grounds, then discuss estimation and finally cover findings from empirical and simulation studies. More specifically, Section 2 describes the forecast combination problem in the context of loss functions that only depend on the forecast error but can have arbitrary shape. Section 3 deals with estimation issues and also covers time-varying and nonlinear combination methods. Section 4 discusses shrinkage combinations while Section 5 covers combinations of interval or density forecasts. Section 6 extracts main conclusions from the empirical literature and conducts a Monte Carlo simulation experiment. Finally Section 7 concludes.

2 The Forecast Combination Problem

Suppose that at time t we are interested in forecasting the future value of some target variable after h periods, whose realization is Y_{t+h}. Since no major new insights arise from the case where Y is multivariate, to simplify the exposition we shall assume that $k = 1$ so $Y_{t+h} \in \mathbb{R}^1$. Forecasts of variables whose realizations at time $t+h$ affect the forecaster's loss are made at time t. We shall refer to h as the forecast horizon. The information set at time t will be denoted by $\mathcal{I}_t$ and we assume that $\mathcal{I}_t$ comprises an $N-$vector of forecasts $\hat{\mathbf{y}}_{t+h,t} = (\hat{y}_{1,t+h,t}, \hat{y}_{2,t+h,t}, ..., \hat{y}_{N,t+h,t})'$ in addition to their histories up to time t and the history of the realizations of the target variable, i.e. $\mathcal{I}_t = \{\hat{\mathbf{y}}_{h+1,1}, \hat{\mathbf{y}}_{t+h,t}, y_1, ..., y_t\}$.[1]

Forecasts do not intrinsically have direct value to decision makers. Rather, they become valuable only as far as they can be used to improve decision makers' actions, which in turn affect their loss or utility. Point forecasts generally provide insufficient information for a decision maker or forecast user who, for example, may be interested in the degree of uncertainty surrounding the forecast. Nevertheless, the vast majority of studies on forecast combinations has dealt with point forecasts so we initially focus on this case. We let $\hat{y}^c_{t+h,t} = g(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_{th})$ be the combined point forecast as a function of the underlying forecasts $\hat{\mathbf{y}}_{t+h,t}$ and the parameters of the combination, $\boldsymbol{\omega}_{th} \in \mathcal{W}_t$, where $\mathcal{W}_t$ is a compact subset of $\mathbb{R}^N$ and $\boldsymbol{\omega}_{th}$ is adapted to $\mathcal{I}_t$. For example, equal weights would give $g(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_{th}) = (1/N)\sum_{j=1}^{N} \hat{\mathbf{y}}_{t+h,t}$. Our choice of notation reflects that we will mostly be thinking of $\boldsymbol{\omega}_{th}$ as combination weights, although

[1] A set of additional information variables, $\mathbf{x}_t$, can easily be included in the problem.

in general the parameters need not have this interpretation.

2.1 Loss Function and Consensus Forecasts

To simplify matters we follow standard practice and assume that the loss function only depends on the forecast error, $e_{t+h,t} = y_{t+h} - \hat{y}^c_{t+h,t}$, i.e. $L = L(e_{t+h})$. The vast majority of work on forecast combinations assume this type of loss, in part because point forecasts are far more common than distribution forecasts and in part because the decision problem underlying the forecast situation is not worked out in detail. However, it should also be acknowledged that this loss function embodies a set of restrictive assumptions on the decision problem, c.f. Granger and Machina (2004) and Elliott and Timmermann (2004). In Section 6 we cover the more general case that combines interval or distribution forecasts.

Letting $e^c_{t+h,t} = y_{t+h} - g(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_{th})$ be the error from the combined forecast, the parameters of the optimal combination, $\boldsymbol{\omega}^*_{th} \in \mathcal{W}_t$, solve the problem

$$\boldsymbol{\omega}^*_{th} = \arg \min_{\boldsymbol{\omega}_{th} \in \mathcal{W}_t} E\left[L\left(e^c_{t+h,t}(\boldsymbol{\omega}_{th})\right) | \hat{\mathbf{y}}_{t+h,t}\right]. \tag{1}$$

Here the expectation is taken over the conditional distribution of $e_{t+h,t}$ given $\mathcal{I}_t$. Elliott and Timmermann (2004) show that, subject to a set of weak technical assumptions on the loss and distribution functions, the combination weights can be found as the solution to the following Taylor series expansion around $\mu_{e_{t+h,t}} = E[e_{t+h,t}|\mathcal{I}_t]$

$$\boldsymbol{\omega}^*_{th} = \min_{\boldsymbol{\omega}_{th} \in \mathcal{W}_t} \left\{ L(\mu_{e_{t+h,t}}) + \frac{1}{2} L''_{\mu_e} E[(e_{t+h,t} - \mu_{e_{t+h,t}})^2 | \mathcal{I}_t] + \sum_{m=3}^{\infty} L^m_{\mu_e} \sum_{i=0}^{m} \frac{1}{i!(m-i)!} E[e^{m-i}_{t+h,t} \mu^i_{e_{t+h,t}} | \mathcal{I}_t] \right\}, \tag{2}$$

where $L^k_{\mu_e} \equiv \partial^k L(e_{t+h,t}) / \partial^k \omega|_{e_{t+h,t} = \mu_{e_{t+h,t}}}$. In general, the entire moment generating function of the error distribution and all higher-order derivatives of the loss function will influence the optimal combination weights which therefore reflect both the shape of the loss function and the forecast error distribution. Furthermore, optimality is established within the assumed family $\hat{y}^c_{t+h,t} = g(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_{th})$.

Oftentimes it is simply assumed that the objective function underlying a combination problem is mean squared error (MSE) loss

$$L(Y_{t+h}, \hat{Y}_{t+h,t}) = \theta(Y_{t+h} - \hat{Y}_{t+h,t})^2, \quad \theta > 0. \tag{3}$$

For this case, the combined or consensus forecast seeks to choose a (possibly time-varying) mapping $g_t(\hat{\mathbf{y}}_{t+h,t};\boldsymbol{\omega}_{th})$ from the N-vector of individual forecasts $\hat{\mathbf{y}}_{t+h,t}$ to the real line, $\widehat{\mathcal{Y}}_{t+h,t} \rightarrow \mathcal{R}$ that best approximates the conditional expectation, $E[Y_{t+h}|\hat{\mathbf{y}}_{t+h,t}]$.[2]

More generally, the forecast combination problem seeks an aggregator that reduces the information in a potentially high-dimensional vector of forecasts, $\hat{\mathbf{y}}_{t+h,t} \in \mathbb{R}^N$, to a lower dimensional summary measure, $C(\hat{\mathbf{y}}_{t+h,t};\boldsymbol{\omega}_c) \in \mathbb{R}^c \subset \mathbb{R}^N$, where $\boldsymbol{\omega}_c$ are the parameters associated with the combination. For example, a decision maker interested in using forecasts to determine how much to invest in a risky asset may want to use information on either the mode, median or mean forecast, but also to consider the degree of dispersion across individual forecasts as a way to measure the uncertainty or 'disagreement' surrounding the forecasts. How low-dimensional the combined forecast should be is not always obvious. Furthermore, outside the MSE framework, it is not trivially true that a scalar aggregator that summarizes all relevant information can always be found. The higher the cross-sectional dispersion, the more important higher order moments become (for a forecast user) relative to the cross-sectional average.

The expansion (2) suggests that the collection of individual forecasts $\hat{\mathbf{y}}_{t+h,t}$ is useful in as far as it can predict any of the conditional moments that a decision maker cares about. Hence, $\hat{y}_{i,t+h,t}$ gets a non-zero weight in the combination if for any moment, $e^m_{t+h,t}$, for which $L^m_{\mu_e} \neq 0$, $\partial E[e^m_{t+h,t}|\mathcal{I}_t]/\partial y_{i,t+h,t} \neq 0$.For example, if the vector of point forecasts can be used to predict the mean, variance, skew and kurtosis but no other moments of the forecast error distribution, then the combined summary measure could be based on summary measures of $\hat{\mathbf{y}}_{t+h,t}$ predicting the first through fourth moments.

As an example of a forecasting service that summarizes different parts of the distribution of forecasts, consider Blue Chip forecasts which provide the following range of forecast measures: "Each forecaster's prediction is published along with the average, or consensus forecast, for each variable. There are also averages of the 10 highest and 10 lowest forecasts for each variable; a median forecast to eliminate the effects of extreme forecasts on the consensus; the number of forecasts raised, lowered, or left unchanged from a month ago; and a

[2]To see this, take expectations of (3) and differentiate with respect to to $g_t(\hat{\mathbf{y}}_{t+h,t};\boldsymbol{\omega}_{th})$ to get $g^*_t(\hat{\mathbf{y}}_{t+h,t};\boldsymbol{\omega}_{th}) = E[Y_{t+h}|\mathcal{I}_t]$.

diffusion index that indicates shifts in sentiment that sometimes occur prior to changes in the consensus forecast."

Two levels of aggregation are thus involved in the combination problem. The first step summarizes/processes individual forecasters' private information to produce individual point forecasts $\hat{y}_{i,t+h,t}$. The only difference to the standard forecasting problem is that the 'input' variables are forecasts from other models or subjective forecasts. This may create a generated regressor problem that can bias the estimated combination weights, although this aspect is typically ignored. It could in part explain why combinations based on estimated weights often do not perform well. A second step aggregates the vector of point forecasts $\hat{\mathbf{y}}_{t+h,t}$ to the consensus measure $C(\hat{\mathbf{y}}_{t+h,t}; \boldsymbol{\omega}_c)$, a dimensionality reduction. Information is lost in both steps. Conversely, the second step is likely to lead to far simpler and more parsimonious forecasting models when compared to a forecast based on the full set of individual forecasts or a "super model" based on individual forecasters' information variables. In general, we would expect information aggregation to increase the bias in the forecast but also to reduce the variance of the forecast error. In as far as possible, the combination should trade off these two components optimally. This is particularly clear under MSE loss, where the objective function equals the squared bias plus the forecast error variance, $E[e^2_{t+h,t}] = E[e_{t+h,t}]^2 + Var(e_{t+h,t})$.

Clemen (1987) demonstrates that an important part of the aggregation of individual forecasts towards an aggregate forecast is an assessment of the dependence among the underlying models' ('experts') forecasts and that a group forecast will generally be less informative than the set of individual forecasts. In fact, group forecasts only provide a sufficient statistic for collections of individual forecasts provided that both the experts and the decision maker agree in their assessments of the dependence among experts. This precludes differences in opinion about the correlation structure among decision makers. Taken to its extreme, this argument suggests that experts should not attempt to aggregate their observed information into a single forecast but should simply report their raw data to the decision maker.

2.2 Construction of a Super Model - pooling information

Let $\mathcal{I}_t^c = \cup_{i=1}^N \mathcal{I}_{it}$ be the union of the forecasters' individual information sets, or the 'super' information set. If $\mathcal{I}_t^c$ were observed, one possibility would be to model the conditional mean of Y_{t+h} as a function of all these variables, i.e.

$$\hat{y}_{t+h,t} = g_s(\mathcal{I}_t^c; \boldsymbol{\theta}_s). \tag{4}$$

Individual forecasts instead take the form $\hat{y}_{i,t+h,t} = g_i(\mathcal{I}_{it}; \boldsymbol{\theta}_i)$. If only the individual forecasts $\hat{y}_{i,t+h,t}$ $(i = 1, .., N)$ are observed, whereas the underlying information sets $\{\mathcal{I}_{it}\}$ are unobserved, the combined forecast would be restricted as follows:

$$\hat{y}_{t+h,t} = g_c(\hat{y}_{1,t+h,t}, ..., \hat{y}_{N,t+h,t}; \boldsymbol{\theta}_c). \tag{5}$$

Normally it would be better to pool all information rather than first filter the information sets through the individual forecasting models, which introduces the usual efficiency loss through the two-stage estimation and also ignores correlations between the underlying information sources. There are several potential problems with pooling of information sets, however. One is—as already mentioned—that individual information sets may not be observable or too costly to combine. Diebold and Pauly (1990, p. 503) remark that "While pooling of forecasts is suboptimal relative to pooling of information sets, it must be recognized that in many forecasting situations, particularly in real time, pooling of information sets is either impossible or prohibitively costly." Furthermore, in cases with many relevant input variables and complicated dynamic and nonlinear effects, constructing a "super model" using the pooled information set, $\mathcal{I}_t^c$, is not likely to provide good forecasts given well-known problems associated with high-dimensional kernel regressions, nearest neighbor regressions or other non-parametric methods. Although individual forecasting models will be biased and may omit important variables, this bias can more than be compensated for by reductions in parameter estimation error in cases where the number of relevant predictor variables is much larger than N. Or, as stated by Yang (2004) p. 179 "A super model of infinite dimension is not very helpful." When the true forecasting model mapping $\mathcal{I}_t^c$ to Y_{t+h} is infinite-dimensional, the model that optimally balances bias and variability will typically

depend on the sample size with a dimension that grows as the sample size increases. The true model may never get recovered although the quality of the approximation increases as the sample size grows.

2.3 Linear Forecast Combinations under MSE Loss

Unless the relation between Y_{t+h} and $\hat{\mathbf{Y}}_{t+h,t}$ is modeled non-parametrically, optimality results must be established within families of parametric combination schemes of the form $Y^c_{t+h,t} = g(\hat{\mathbf{Y}}_{t+h,t}; \boldsymbol{\theta}_t)$. The general class of combination schemes in (1) comprises non-linear as well as time-varying combination methods. We shall return to these but for now concentrate on the family (or subset) of linear combinations, $\mathcal{W}^l \subset \mathcal{W}$, which are more commonly used.[3] To this end we choose weights, $\boldsymbol{\omega}_{t,h} = (\omega_{1t+h,t}, ..., \omega_{N,t+h,t})'$ to produce a combined forecast of the form

$$\hat{y}^c_{t+h,t} = \boldsymbol{\omega}'_{t,h}\hat{\mathbf{y}}_{t+h,t}. \tag{6}$$

Without risk of confusion we have dropped the subscript that refers to the forecast horizon, h. While in general there is no closed-form solution to (1), one can get analytical results by imposing distributional restrictions or restrictions on the loss function. Under MSE loss, the combination weights are easy to characterize in population and only depend on the first two conditional moments of the joint distribution of Y and $\hat{\mathbf{Y}}$,

$$\begin{pmatrix} Y_{t+h} \\ \hat{\mathbf{Y}}_{t+h,t} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{yth} \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}th} \end{pmatrix} \begin{pmatrix} \sigma^2_{yth} & \boldsymbol{\sigma}'_{y\hat{\mathbf{y}}th} \\ \boldsymbol{\sigma}_{y\hat{\mathbf{y}}th} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th} \end{pmatrix} \right). \tag{7}$$

Minimizing $E[e^2_{t+h,t}] = E[(Y_{t+h} - \boldsymbol{\omega}'_{t+h,t}\hat{\mathbf{y}}_{t+h,t})^2]$, we have

$$\boldsymbol{\omega}^*_{t,h} = \arg \min_{\boldsymbol{\omega}_{t,h} \in \mathcal{W}^l} \left((\mu_{yth} - \boldsymbol{\omega}'_{t,h}\boldsymbol{\mu}_{th})^2 + \sigma^2_{yth} + \boldsymbol{\omega}'_{t,h}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\omega}_{t,h} - 2\boldsymbol{\omega}'_{t,h}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th} \right).$$

This yields the first order condition

$$\frac{\partial E[e^2_{t+h,t}]}{\partial \boldsymbol{\omega}_{t,h}} = -2(\mu_{yth} - \boldsymbol{\omega}'_{t,h}\boldsymbol{\mu}_{th})\boldsymbol{\mu}_{th} + 2\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\omega}_{t,h} - 2\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th} = 0,$$

[3] This, of course, does not rule out that the *estimated* weights vary over time as will be the case when the weights are updated recursively as more data becomes available.

with solution−assuming that $\mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}$ is invertible−

$$\boldsymbol{\omega}_{th}^{*} = (\boldsymbol{\mu}_{th}\boldsymbol{\mu}_{th}' + \mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th})^{-1}(\boldsymbol{\mu}_{th}\mu_{yth} + \boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}). \tag{8}$$

This solution is optimal in population whenever Y_{t+h} and $\hat{\mathbf{Y}}_{t+h,t}$ are joint Gaussian since in this case the conditional expectation $E[Y_{t+h}|\hat{\mathbf{Y}}_{t+h,t}]$ will be linear in $\hat{\mathbf{Y}}_{t+h,t}$. A constant can trivially be included as one of the forecasts so that the combination scheme allows for an intercept term, a strategy recommended (under MSE loss) by Granger and Ramanathan (1984) and (for a variety of loss functions) by Elliott and Timmermann (2004). Assuming that the first forecast is in fact a constant, the optimal (population) values of the constant and the combination weights, ω_{0th}^{*} and $\boldsymbol{\omega}_{th}^{*}$, simplify as follows

$$\begin{aligned} \omega_{0th}^{*} &= \mu_{yth} - \boldsymbol{\omega}_{th}^{*\prime}\boldsymbol{\mu}_{th}, \\ \boldsymbol{\omega}_{th}^{*} &= \mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}^{-1}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}. \end{aligned} \tag{9}$$

These weights depend on the full conditional covariance matrix of forecast errors, $\mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}$. In general the weights have an intuitive interpretation and tend to be larger for more accurate forecasts that are less strongly correlated with other forecasts.

In the following we explore some interesting special cases to demonstrate the determinants of gains from forecast combination.

2.3.1 Diversification Gains

Under quadratic loss it is easy to illustrate the population gains from different forecast combination schemes. This is an important task since, as argued by Winkler (1989, p. 607) "The better we understand which sets of underlying assumptions are associated with which combining rules, the more effective we will be at matching combining rules to forecasting situations." To this end we consider the simple combination of two forecasts that give rise to errors $e_1 = Y - \hat{Y}_1$ and $e_2 = Y - \hat{Y}_2$. Without risk of confusion we have dropped the time subscripts. Assuming that the individual forecast errors are unbiased, we have $e_1 \sim (0, \sigma_1^2), e_2 \sim (0, \sigma_2^2)$ where $\sigma_1^2 = var(e_1), \sigma_2^2 = var(e_2), \sigma_{12} = \rho_{12}\sigma_1\sigma_2$ is the covariance between e_1 and e_2 and ρ_{12} is their correlation. Suppose that the combination weights are

restricted to sum to one, with weights $(\omega, 1-\omega)$ on the first and second forecast. The forecast error from the combination takes the form

$$e^c = \omega e_1 + (1-\omega)e_2 \tag{10}$$

with variance

$$\sigma_c^2(\omega) = \omega^2\sigma_1^2 + (1-\omega)^2\sigma_2^2 + 2\omega(1-\omega)\sigma_{12}. \tag{11}$$

Differentiating with respect to ω and solving the first order condition, we have

$$\begin{aligned} \omega^* &= \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \\ 1-\omega^* &= \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}. \end{aligned} \tag{12}$$

A greater weight is assigned to models with more precise forecasts. A negative weight on a forecast clearly does not mean that it has no value. In fact when $\rho_{12} > \sigma_2/\sigma_1$ the combination weights are not convex and one weight will exceed unity, the other being negative, c.f. Bunn (1985).

Inserting ω^* into (11), we get the expected squared loss associated with the optimal weights:

$$\sigma_c^2(\omega^*) = \frac{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}. \tag{13}$$

It can easily be verified that $\sigma_c^2(\omega^*) \leq \min(\sigma_1^2, \sigma_2^2)$. In fact, the diversification gain will only be zero in the following special cases (i) σ_1 or σ_2 equal to zero; (ii) $\sigma_1 = \sigma_2$ and $\rho_{12} = 1$; or (iii) $\rho_{12} = \sigma_1/\sigma_2$, c.f. Bunn (1985).

It is interesting to compare the variance of the forecast error from the optimal combination to the variance of the combination scheme that weights the forecasts inversely to their relative mean squared error (MSE) values and hence ignores any correlation between the forecast errors:

$$\omega_{inv} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad 1-\omega_{inv} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \tag{14}$$

These weights result in a forecast error variance

$$\sigma_{inv}^2 = \frac{\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2 + 2\rho_{12}\sigma_1\sigma_2)}{(\sigma_1^2 + \sigma_2^2)^2}. \tag{15}$$

After some algebra we can derive the ratio of the forecast error variance under this scheme relative to its value under the optimal weights, $\sigma_c^2(\omega^*)$:

$$\frac{\sigma_{inv}^2}{\sigma_c^2(\omega^*)} = \left(\frac{1}{1-\rho_{12}^2}\right)\left(1-\left(\frac{2\sigma_{12}}{\sigma_1^2+\sigma_2^2}\right)^2\right). \tag{16}$$

If $\sigma_1 \neq \sigma_2$, this exceeds unity unless $\rho_{12} = 0$. When $\sigma_1 = \sigma_2$, this ratio is always unity irrespective of the value of ρ_{12} as in this case $\omega_{inv} = \omega^* = 1/2$. Thus equal weights are optimal when combining two forecasts provided that the two forecast error variances are identical, irrespective of the correlation between the two forecast errors.

Another interesting benchmark is the equal-weighted combination $\hat{y}^{ew} = (1/2)(\hat{y}_1 + \hat{y}_2)$. Under these weights the variance of the forecast error is

$$\sigma_{ew}^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1\sigma_2\rho_{12} \tag{17}$$

so the ratio $\sigma_{ew}^2/\sigma_c^2(\omega^*)$ becomes:

$$\frac{\sigma_{ew}^2}{\sigma_c^2(\omega^*)} = \left(\frac{(\sigma_1^2+\sigma_2^2)^2 - 4\sigma_{12}^2}{4\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}\right), \tag{18}$$

which in general exceeds unity unless $\sigma_1 = \sigma_2$.

Finally, as a measure of the diversification gain obtained from combining the two forecasts it is natural to compare $\sigma_c^2(\omega^*)$ to $\min(\sigma_1^2, \sigma_2^2)$. Suppose that $\sigma_1 > \sigma_2$ and define $\kappa = \sigma_2/\sigma_1$ so that $\kappa < 1$. We then have

$$\frac{\sigma_c^2(\omega^*)}{\sigma_2^2} = \frac{1-\rho_{12}^2}{1+\kappa^2-2\rho_{12}\kappa}. \tag{19}$$

Figure 1 shows this expression graphically as a function of ρ_{12} and κ. The (relative) diversification gain is a complicated function of the correlation between the two forecast errors, ρ_{12}, and the relative variance of the forecast errors, κ. In fact, the sign of the derivative of the efficiency gain with respect to either κ or ρ_{12} changes even for reasonable parameter values. Differentiating (19) with respect to ρ_{12}, we have

$$\partial\left(\frac{\sigma_c^2(\omega^*)}{\sigma_2^2}\right)/\partial\rho_{12} \propto \kappa\rho_{12}^2 - (1+\kappa^2)\rho_{12} + \kappa.$$

This is a second order polynomial in ρ_{12} with roots (assuming $\kappa < 1$)

$$\frac{1+\kappa^2 \pm (1-\kappa^2)}{2\kappa} = (\kappa; 1/\kappa).$$

Only when $\kappa = 1$ (so $\sigma_1^2 = \sigma_2^2$) does it follow that the efficiency gain will be an increasing function of ρ_{12} - otherwise it will change sign, being positive on the interval $[-1; \kappa]$ and negative on $[\kappa; 1]$ as can be seen from Figure 1. This figure shows that diversification is more effective (in the sense that it results in the largest reduction in the efficiency ratio for a given change in ρ_{12}) when $\kappa = 1$.

2.3.2 Effect of Bias in individual forecasts

Problems can arise for forecast combinations when one or more of the individual forecasts is biased, the combination weights are constrained to sum to unity and an intercept is omitted from the combination. Min and Zellner (1993) illustrate how bias in one or more of the forecasts along with a constraint that the weights add up to unity can lead to suboptimality of combinations. Let $y - \hat{y}_1 = e_1 \sim (0, \sigma^2)$ and $y - \hat{y}_2 = e_2 \sim (\mu_2, \sigma^2)$, $cov(e_1, e_2) = \sigma_{12} = \rho_{12}\sigma^2$, so $\hat{y}_1$ is unbiased while $\hat{y}_2$ is biased with a bias equal to μ_2. Then the MSE of $\hat{y}_1$ is σ^2, while the MSE of $\hat{y}_2$ is $\sigma^2 + \mu_2^2$. The MSE of the combined forecast $\hat{y}_c = \omega\hat{y}_1 + (1-\omega)\hat{y}_2$ relative to that of the best forecast ($\hat{y}_1$) is

$$MSE(\hat{y}_c) - MSE(\hat{y}_1) = (1-\omega)\sigma^2 \left((1-\omega) MSE(\hat{y}_c) \left(\frac{\mu_2}{\sigma}\right)^2 - MSE(\hat{y}_1) - 2\omega(1-\rho_{12}) \right),$$

so $MSE(\hat{y}_c) > MSE(\hat{y}_1)$ if

$$\left(\frac{\mu_2}{\sigma}\right)^2 > \frac{2\omega(1-\rho_{12})}{1-\omega}.$$

This condition holds if $\rho_{12} = 1$. Furthermore, the larger the bias, the more likely it is that the combination will not dominate the first forecast.

2.4 Optimality of Equal weights - general case

Equal weight occupy a special place in the forecast combination literature. They are frequently either imposed on the combination scheme or used as a point towards which the unconstrained combination weights are shrunk. Given their special role, it is worthwhile establishing more general conditions under which they are optimal in a population sense. This sets a benchmark that proves helpful in understanding their good finite-sample performance in simulations and experiments with actual data.

Let $\mathbf{\Sigma}_e = E[\mathbf{ee}']$ be the covariance matrix of the forecast errors. Again we drop time subscripts without any risk of confusion. From (7) the vector of forecast errors $\mathbf{e} = \boldsymbol{\iota} Y - \hat{\mathbf{Y}}$ (where $\boldsymbol{\iota}$ is an $N \times 1$ column vector of ones) has second moment

$$\begin{aligned} \mathbf{\Sigma}_e &= E[Y^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \hat{\mathbf{Y}}\hat{\mathbf{Y}}' - 2Y\boldsymbol{\iota}\hat{\mathbf{Y}}'] \\ &= (\sigma_y^2 + \mu_y^2)\boldsymbol{\iota}\boldsymbol{\iota}' + \boldsymbol{\mu}\boldsymbol{\mu}' + \mathbf{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - 2\boldsymbol{\iota}\boldsymbol{\sigma}'_{y\hat{\mathbf{y}}} - 2\mu_y \boldsymbol{\iota}\boldsymbol{\mu}'. \end{aligned} \tag{20}$$

Consider minimizing the expected forecast error variance subject to the constraint that the weights add up to one:

$$\begin{aligned} &\min \boldsymbol{\omega}' \mathbf{\Sigma}_e \boldsymbol{\omega} \\ &s.t.\ \boldsymbol{\omega}'\boldsymbol{\iota} = 1. \end{aligned} \tag{21}$$

The constraint ensures unbiasedness of the combined forecast provided that $\boldsymbol{\mu} = \mu_y \boldsymbol{\iota}$ so that

$$\mu_y^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \boldsymbol{\mu}\boldsymbol{\mu}' - 2\mu_y \boldsymbol{\iota}\boldsymbol{\mu}' = 0.$$

The Lagrangian associated with (21) is

$$\mathcal{L} = \boldsymbol{\omega}' \mathbf{\Sigma}_e \boldsymbol{\omega} - \lambda(\boldsymbol{\omega}'\boldsymbol{\iota} - 1)$$

which yields the first order condition

$$\mathbf{\Sigma}_e \boldsymbol{\omega} = \frac{\lambda}{2}\boldsymbol{\iota}. \tag{22}$$

Assuming that $\mathbf{\Sigma}_e$ is invertible, after pre-multiplying by $\mathbf{\Sigma}_e^{-1}\boldsymbol{\iota}'$ and recalling that $\boldsymbol{\iota}'\boldsymbol{\omega} = 1$ we get

$$\frac{\lambda}{2} = \left(\boldsymbol{\iota}' \mathbf{\Sigma}_e^{-1} \boldsymbol{\iota}\right)^{-1}.$$

Inserting this in (22) we have the frequently cited formula for the optimal weights:

$$\boldsymbol{\omega}^* = (\boldsymbol{\iota}' \mathbf{\Sigma}_e^{-1} \boldsymbol{\iota})^{-1} \mathbf{\Sigma}_e^{-1} \boldsymbol{\iota}. \tag{23}$$

Now suppose that the forecast errors have the same variance, σ^2, and correlation, ρ. Then we have

$$\begin{aligned} \mathbf{\Sigma}_e^{-1} &= \frac{1}{\sigma^2(1-\rho)}\left(\mathbf{I} - \frac{\rho}{1+(N-1)\rho}\boldsymbol{\iota}\boldsymbol{\iota}'\right) \\ &= \frac{1}{\sigma^2(1-\rho)(1+(N-1)\rho)}\left((1+(N-1)\rho)\mathbf{I} - \rho\boldsymbol{\iota}\boldsymbol{\iota}'\right), \end{aligned}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Inserting this in (23) we have

$$\begin{aligned} \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\iota} &= \frac{\boldsymbol{\iota}}{\sigma^2(1+(N-1)\rho)} \\ (\boldsymbol{\iota}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{\iota})^{-1} &= \frac{\sigma^2(1+(N-1)\rho)}{N}, \end{aligned}$$

so

$$\boldsymbol{\omega}^* = \left(\frac{1}{N}\right)\boldsymbol{\iota}. \tag{24}$$

Hence equal-weights are optimal in situations with an arbitrary number of forecasts when the individual forecast errors have the same variance and identical pair-wise correlations. Notice that the property that the weights add up to unity only follows as a result of imposing the constraint $\boldsymbol{\iota}'\boldsymbol{\omega} = 1$ and will not otherwise hold more generally.

2.5 Optimal Combinations under Asymmetric Loss

Recent work has seen considerable interest in the effect of asymmetric loss on optimal predictions, c.f., inter alia, Christoffersen and Diebold (1997), Granger and Pesaran (2000) and Patton and Timmermann (2004). These papers show that the standard properties of an optimal forecast under MSE loss−lack of bias, absence of serial correlation in the forecast error at the single-period forecast horizon and increasing forecast error variance as the horizon grows−cease to hold under asymmetric loss. It is therefore not surprising that asymmetric loss also affects combination weights. To illustrate the significance of the shape of the loss function for the optimal combination weights, consider linex loss. The linex loss function is convenient to use since it allows us to characterize the optimal forecast analytically. It takes the form

$$L(e_{t+h,t}) = \exp(ae_{t+h,t}) - ae_{t+h,t} + 1, \tag{25}$$

where a is a scalar that controls the aversion towards either positive ($a > 0$) or negative ($a < 0$) forecast errors and $e_{t+h,t} = (y_{t+h} - \omega_{0th} - \boldsymbol{\omega}'_{th}\widehat{\mathbf{y}}_{t+h,t})$. First, suppose that the target variable and forecast are joint Gaussian with moments given in (7). Using the well-known result that if $X \sim N(\mu, \sigma^2)$, then $E[e^x] = \exp(\mu + \sigma^2/2)$, the optimal combination weights

$(\omega^*_{0th}, \boldsymbol{\omega}^*_{th})$ which minimize the expected loss $E[L(e_{t+h,t})|\mathcal{I}_t]$, solve

$$\min_{\omega_{0th},\boldsymbol{\omega}_{th}} \Big\{ \exp(a(\mu_{yth} - \omega_{0th} - \boldsymbol{\omega}'_{th}\boldsymbol{\mu}_{th}) + \frac{a^2}{2}(\sigma^2_{yth} + \boldsymbol{\omega}'_{th}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\omega}_{th} - 2\boldsymbol{\omega}'_{th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th})) \\ -a(\mu_{yth} - \omega_{0th} - \boldsymbol{\omega}'_{th}\boldsymbol{\mu}_{th}) \Big\}.$$

Taking derivatives we get the first order conditions

$$\begin{aligned} \exp(a(\mu_{yth} - \omega_{0th} - \boldsymbol{\omega}'_{th}\boldsymbol{\mu}_{th}) + \frac{a^2}{2}(\sigma^2_{yth} + \boldsymbol{\omega}'_{th}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\omega}_{th} - 2\boldsymbol{\omega}'_{th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th})) &= 1 \\ \exp(.)(-a\boldsymbol{\mu}_{th} + \frac{a^2}{2}(2\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\omega}_{th} - 2\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th})) + a\boldsymbol{\mu}_{th} &= 0, \end{aligned} \tag{26}$$

where exp(.) is the exponential term from the first equation and hence equals one. Using this we can see that $\boldsymbol{\omega}^*_{th} = \boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}$ which when inserted in the first equation gives the optimal solution

$$\begin{aligned} \omega_{0th} &= \mu_{yth} - \boldsymbol{\omega}^{*\prime}_{th}\boldsymbol{\mu}_{th} + \frac{a}{2}(\sigma^2_{yth} - \boldsymbol{\omega}^{*\prime}_{th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}), \\ \boldsymbol{\omega}^*_{th} &= \boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}. \end{aligned} \tag{27}$$

Notice that the optimal combination weights, $\boldsymbol{\omega}^*_{th}$, are unchanged from the case with MSE loss, (9), while the intercept accounts for the shape of the loss function and depends on the parameter a. In fact, the optimal combination will have a bias, $\frac{a}{2}(\sigma^2_{yth} - \boldsymbol{\omega}^{*\prime}_{th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th})$, that reflects the dispersion of the forecast error evaluated at the optimal combination weights.

Next, suppose that we allow for a non-Gaussian forecast error distribution. We do so by assuming that the joint distribution of $(Y_{t+h}\ \hat{\mathbf{Y}}'_{t+h,t})'$ is a mixture of two Gaussian distributions driven by a state variable, S_{t+h}, which can take two values, i.e. $s_{t+h} = 1$ or $s_{t+h} = 2$ so that

$$\begin{pmatrix} Y_{t+h} \\ \hat{\mathbf{Y}}_{t+h,t} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{ys_{t+h}} \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}s_{t+h}} \end{pmatrix}, \begin{pmatrix} \sigma^2_{ys_{t+h}} & \boldsymbol{\sigma}'_{y\hat{\mathbf{y}}s_{t+h}} \\ \boldsymbol{\sigma}_{y\hat{\mathbf{y}}s_{t+h}} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}s_{t+h}} \end{pmatrix} \right). \tag{28}$$

Furthermore, suppose that $P(S_{t+h} = 1) = p$, while $P(S_{t+h} = 2) = 1 - p$. The two regimes could correspond to recession and expansion states for the economy (Hamilton (1989)) or bull and bear states for financial markets, c.f. Guidolin and Timmermann (2003).

Under this model,

$$\begin{aligned} e_{t+h,t} &= Y_{t+h} - \omega_{0th} - \boldsymbol{\omega}'_{th}\hat{\mathbf{Y}}_{t+h,t} \\ &\sim N\left(\mu_{ys_{t+h}} - \omega_{0th} - \boldsymbol{\omega}'_{th}\boldsymbol{\mu}_{\hat{\mathbf{y}}s_{t+h}}, \sigma^2_{ys_{t+h}} + \boldsymbol{\omega}'_{th}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}s_{t+h}}\boldsymbol{\omega}_{th} - 2\boldsymbol{\omega}'_{th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}s_{t+h}}\right). \end{aligned}$$

Dropping time subscripts, the expected loss under this distribution, $E[L(e_{t+h,t})|\widehat{\mathbf{y}}_{t+h,t}]$, is proportional to

$$p\left\{\exp(a(\mu_{y1}-\omega_0-\boldsymbol{\omega}'\boldsymbol{\mu}_{\widehat{\mathbf{y}}1})+\frac{a^2}{2}(\sigma^2_{y1}+\boldsymbol{\omega}'\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1}\boldsymbol{\omega}-2\boldsymbol{\omega}'\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1}))-a(\mu_{y1}-\omega_0-\boldsymbol{\omega}'\boldsymbol{\mu}_{\widehat{\mathbf{y}}1})\right\}$$
$$+(1-p)\left\{\exp(a(\mu_{y2}-\omega_0-\boldsymbol{\omega}'\boldsymbol{\mu}_{\widehat{\mathbf{y}}2})+\frac{a^2}{2}(\sigma^2_{y2}+\boldsymbol{\omega}'\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2}\boldsymbol{\omega}-2\boldsymbol{\omega}'\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}2}))-a(\mu_{y2}-\omega_0-\boldsymbol{\omega}'\boldsymbol{\mu}_{\widehat{\mathbf{y}}2})\right\}.$$

Taking derivatives, we get the following first order conditions for ω_0 and $\boldsymbol{\omega}$

$$p(\exp(\xi_1)-1)+(1-p)(\exp(\xi_2)-1)=0,$$

$$p\left(\exp(\xi_1)(-\boldsymbol{\mu}_{\widehat{\mathbf{y}}1}+\frac{a}{2}(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1}\boldsymbol{\omega}-\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1}))+\boldsymbol{\mu}_{\widehat{\mathbf{y}}1}\right)+$$
$$(1-p)\left(\exp(\xi_2)(-\boldsymbol{\mu}_{\widehat{\mathbf{y}}2}+\frac{a}{2}(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2}\boldsymbol{\omega}-\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}2}))+\boldsymbol{\mu}_{\widehat{\mathbf{y}}2}\right)=0,$$

where $\xi_{s_{t+1}}=a(\mu_{ys_{t+1}}-\omega_0-\boldsymbol{\omega}'\boldsymbol{\mu}_{\widehat{\mathbf{y}}s_{t+1}})+\frac{a^2}{2}(\sigma^2_{ys_{t+1}}+\boldsymbol{\omega}'\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}s_{t+1}}\boldsymbol{\omega}-2\boldsymbol{\omega}'\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}s_{t+1}})$. In general this gives a set of $N+1$ highly non-linear equations in ω_0 and $\boldsymbol{\omega}$. The exception is when $\boldsymbol{\mu}_{\widehat{\mathbf{y}}1}=\boldsymbol{\mu}_{\widehat{\mathbf{y}}2}$, in which case (using the first order condition for ω_0) the first order condition for $\boldsymbol{\omega}$ simplifies to

$$p\exp(\xi_1)(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1}\boldsymbol{\omega}-\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1}))+(1-p)\exp(\xi_2)(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2}\boldsymbol{\omega}-\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}2}))=0.$$

When $\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2}=\varphi\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1}$ and $\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}2}=\varphi\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1}$, the solution to this equation again corresponds to the optimal weights for the MSE loss function, (9):

$$\boldsymbol{\omega}^*=\boldsymbol{\Sigma}^{-1}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1}\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1}. \tag{29}$$

This restriction of course represents a special case and ensures that the joint distribution of $(Y_{t+h},\hat{\mathbf{Y}}_{t+h,t})$ is elliptically symmetric−a class of distributions that encompasses the multivariate Gaussian. We just showed a special case of the more general result showed by Elliott and Timmermann (2004): if the joint distribution of $(Y_{t+h}\ \hat{\mathbf{Y}}'_{t+h,t})'$ is elliptically symmetric and the expected loss can be written as a function of the mean and variance of the forecast error, μ_e and σ^2_e, i.e., $E[L(e_t)]=g(\mu_e,\sigma^2_e)$, then the optimal forecast combination weights take the form (29) and hence do not depend on the shape of the loss function (other than for certain technical conditions), while conversely the constant (ω_0) reflects this shape. Thus,

under fairly general conditions on the loss functions, a forecast enters into the optimal forecast combination with a non-zero weight if and only if its optimal weight under MSE loss is non-zero. Conversely, if the conditions ensuring the elliptical symmetry fail to hold, then it is quite possible that a forecast may have a non-zero weight under functions other than MSE loss but not under MSE loss and vice versa. The latter case is likely to be most relevant empirically since studies using regime switching models often find that although the mean parameters may be constrained to be identical across the regimes, the variance-covariance parameters tend to be very different across regimes, c.f., e.g. Guidolin and Timmermann (2003).

This example can be used to demonstrate that a forecast that does not work most of the time (in the sense that it is uncorrelated with the outcome variable) but does so only a small part of the time when other forecasting variables happen to break down is still valuable. We set all mean parameters equal to one, $\mu_{y1} = \mu_{y2} = 1$, $\boldsymbol{\mu}_{\widehat{\mathbf{y}}1} = \boldsymbol{\mu}_{\widehat{\mathbf{y}}2} = \boldsymbol{\iota}$, so bias can be ignored, while the variance-covariance parameters are chosen as follows

$$\begin{aligned}
\sigma_{y1} &= 3;\ \sigma_{y2} = 1, \\
\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1} &= 0.8 \times \sigma_{y1}^2 \times \mathbf{I}\ \ ;\ \boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2} = 0.5 \times \sigma_{y2}^2 \times \mathbf{I} \\
\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}1} &= \sigma_{y1} \times \sqrt{diag(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}1})} \odot \begin{pmatrix} 0.9 \\ 0.2 \end{pmatrix}, \\
\boldsymbol{\sigma}_{y\widehat{\mathbf{y}}2} &= \sigma_{y2} \times \sqrt{diag(\boldsymbol{\Sigma}_{\widehat{\mathbf{y}}\widehat{\mathbf{y}}2})} \odot \begin{pmatrix} 0.0 \\ 0.8 \end{pmatrix},
\end{aligned}$$

where $\odot$ is the Hadamard or element by element multiplication operator.

In Table 1 we show the optimal weight on the two forecasts as a function of p for two different values of a, namely $a = 1$, corresponding to strongly asymmetric loss, and $a = 0.1$, representing less asymmetric loss. When $p = 0.05$ and $a = 1$, so there is only a five percent chance that the process is in state 1, the optimal weight on model 1 is 35%. This is lowered to only 8% when the asymmetry parameter is reduced to $a = 0.1$. Hence the low probability event has a greater effect on the optimal combination weights the higher the degree of asymmetry in the loss function and the higher the variability of such events.

Table 1: Optimal weights under asymmetric loss

$a = 1$			$a = 0.1$		
p	ω_1^*	ω_2^*	p	ω_1^*	ω_2^*
0.05	0.346	0.324	0.05	0.081	0.365
0.10	0.416	0.314	0.10	0.156	0.353
0.25	0.525	0.297	0.25	0.354	0.323
0.50	0.636	0.280	0.50	0.620	0.283
0.75	0.744	0.264	0.75	0.831	0.250
0.90	0.842	0.249	0.90	0.940	0.234

This example can also be used to demonstrate why forecast combinations may work when the underlying predictors are generated under different loss functions. Suppose that two forecasters have linex loss with parameters $a_1 > 0$ and $a_2 < 0$ and suppose that both have access to the same information set and use the same model to forecast the mean and variance of Y, $\hat{\mu}_y$, $\hat{\sigma}_y^2$. Their forecasts are then computed as (c.f., Christoffersen and Diebold (1997))

$$\begin{aligned} \hat{y}_{1,t+1,t} &= \hat{\mu}_y + \frac{a_1}{2}\hat{\sigma}_y^2, \\ \hat{y}_{2,t+1,t} &= \hat{\mu}_y + \frac{a_2}{2}\hat{\sigma}_y^2. \end{aligned}$$

Each forecast includes an (optimal) bias whose magnitude is time-varying. For a forecast user with symmetric loss, neither of these forecasts is particularly useful as each is biased. Furthermore, the bias cannot simply be taken out by including a constant in the forecast combination regression since the bias is time-varying. However, in this simple case, an exact linear combination of the two forecasts that is unbiased exists:

$$\begin{aligned} \hat{y}_{t+1,t}^c &= \omega\hat{y}_{1,t+1,t} + (1-\omega)\hat{y}_{2,t+1,t} \\ \omega &= \frac{-a_2}{a_1 - a_2}. \end{aligned}$$

Of course this is a special case, but it nevertheless does show how bias in individual forecasts can wash out in a forecast combination.

2.6 Combining as a Hedge against Non-stationarities

Hendry and Clements (2002) argue that combinations may work because they provide insurance against what they refer to as extraneous (deterministic) structural breaks. In simulations they provide supporting evidence that simple combinations can work well under breaks in the explanatory variables. Hendry and Clements consider a wide array of designs for the break and find that combinations work well under a shift in the intercept of a single variable in the data generating process or when two or more positively correlated predictor variables are subject to shifts in opposite directions - in which case forecast combinations can be expected to lead to even larger reductions in the MSE. Their analysis considers the case where a break occurs after the estimation period so that it does not affect the parameter estimates of the individual forecasting models. They establish analytically conditions on the size of the post-sample break that ensure that an equal-weighted combination out-performs the individual forecasts.

Winkler (1989) argues (p. 606) that "... in many situations there is no such thing as a 'true' model for forecasting purposes. The world around us is continually changing, with new uncertainties replacing old ones."

In support of the interpretation that structural breaks or model instability may explain the good average or overall performance of forecast combination methods, Stock and Watson (2003) report that the performance of combined forecasts tends to be far more stable than that of the individual constituent forecasts entering in the combinations. Interestingly, however, many of the combination methods that attempt to build in time-variations in the combination weights (either in the form of discounting of past performance or time-varying parameters) have generally not proved to be successful, although there have been exceptions.

It is easy to construct examples of specific forms of non-stationarities in the underlying data generating process for which simple combinations work better than the single best forecast. Aiolfi and Timmermann (2004) study the following simple model for changes or

shifts in the data generating process:

$$\begin{aligned} Y_t &= S_t F_{1t} + (1 - S_t) F_{2t} + \varepsilon_{yt}, \\ \hat{Y}_{1t} &= F_{1t} + \varepsilon_{1t}, \\ \hat{Y}_{2t} &= F_{2t} + \varepsilon_{2t}. \end{aligned} \tag{30}$$

All variables are assumed to be Gaussian with factors $F_{1t} \sim N(\mu_1, \sigma^2_{F_1})$, $F_{2t} \sim N(\mu_2, \sigma^2_{F_2})$ and innovations $\varepsilon_{yt} \sim N(0, \sigma^2_{\varepsilon_Y})$, $\varepsilon_{1t} \sim N(0, \sigma^2_{\varepsilon_1})$, $\varepsilon_{2t} \sim N(0, \sigma^2_{\varepsilon_2})$. All the innovations are mutually uncorrelated and uncorrelated with the factors, while $Cov(F_{1t}, F_{2t}) = \sigma_{F_1 F_2}$. They further assume that $P(S_t = 1) = p$, $P(S_t = 0) = 1 - p$. Define the population projection coefficient of Y_t on $\hat{Y}_{1t}$ as β_1 and the population projection coefficient of Y_t on $\hat{Y}_{2t}$ as β_2, so that

$$\begin{aligned} \beta_1 &= \frac{p\sigma^2_{F_1} + (1-p)\sigma_{F_1 F_2}}{\sigma^2_{F_1} + \sigma^2_{\varepsilon_1}}, \\ \beta_2 &= \frac{(1-p)\sigma^2_{F_2} + p\sigma^2_{F_1}}{\sigma^2_{F_2} + \sigma^2_{\varepsilon_2}}. \end{aligned}$$

The first and second moments of the forecast errors $e_{it} = Y_t - \hat{Y}_{it}$, can then be characterized as follows::

Conditional on $S_t = 1$:

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim N\left(\begin{pmatrix} (1-\beta_1)\mu_1 \\ \mu_1 - \beta_2\mu_2 \end{pmatrix}, \begin{pmatrix} (1-\beta_1)^2\sigma^2_{F_1} + \beta_1^2\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_y} & (1-\beta_1)\sigma^2_{F_1} + \sigma^2_{\varepsilon y} \\ (1-\beta_1)\sigma^2_{F_1} + \sigma^2_{\varepsilon_Y} & \sigma^2_{F_1} + \beta_2^2\sigma^2_{F_2} + \beta_2^2\sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_y} \end{pmatrix} \right)$$

Conditional on $S_t = 0$:

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_2 - \beta_1\mu_1 \\ (1-\beta_2)\mu_2 \end{pmatrix}, \begin{pmatrix} \beta_1^2\sigma^2_{F_1} + \sigma^2_{F_2} + \beta_1^2\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_Y} & (1-\beta_2)\sigma^2_{F_2} + \sigma^2_{\varepsilon_Y} \\ (1-\beta_2)\sigma^2_{F_2} + \sigma^2_{\varepsilon_Y} & (1-\beta_2)^2\sigma^2_{F_2} + \beta_2^2\sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_Y} \end{pmatrix} \right)$$

Under the joint distribution of $(Y_t, \hat{Y}_{1t}, \hat{Y}_{2t})$ in (30), Aiolfi and Timmermann (2004) show that the population MSE of the equal-weighted combined forecast will be lower than the population MSE of the best model provided that the following condition holds:

$$\frac{1}{3}\left(\frac{p}{1-p}\right)^2 \frac{(1+\psi_2)}{(1+\psi_1)} < \frac{\sigma^2_{F_2}}{\sigma^2_{F_1}} < 3\left(\frac{p}{1-p}\right)^2 \frac{(1+\psi_2)}{(1+\psi_1)}.$$

Here $\psi_1 = \sigma^2_{\varepsilon_1}/\sigma^2_{F_1}, \psi_2 = \sigma^2_{\varepsilon_2}/\sigma^2_{F_2}$ are the noise-to-signal ratios for forecasts one and two, respectively. Hence if $p = 1 - p = 1/2$ and $\psi_1 = \psi_2$ we get the inequality

$$\frac{1}{3} < \frac{\sigma^2_{F_2}}{\sigma^2_{F_1}} < 3,$$

suggesting that equal-weighted combinations will provide a hedge against 'breaks' for a wide range of values of the relative factor variance. How good an approximation this model is to actual data can be debated, but regime shifts have been widely documented for first and second moments of, *inter alia*, output growth, stock and bond returns, interest rates and exchange rates.

Conversely, when combination weights have to be estimated, instability in the data generating process may cause underperformance relative to that of a good individual model. Hence we can construct examples where combination is the dominant strategy in the absence of breaks or other forms of non-stationarities, but becomes inferior in the presence of breaks. Note that this may happen if the conditional distribution of the target variable given a particular forecast is stationary, whereas the correlations between the forecasts changes in which case the combination weights will change but the individual models' performance will remain the same.

3 Estimation

Combining two or more forecasts, while appealing in theory, has the disadvantage over using a single forecast that it introduces parameter estimation error in cases where the combination weights need to be estimated. This is an important point - so much so, that seemingly suboptimal combination schemes such as equal-weighting have widely been found to dominate combination methods that would be optimal in the absence of parameter estimation errors. Finite-sample errors in the estimates of the combination weights can lead to poor performance of combination schemes that dominate in large samples, in part due to the prevalence of strong multicollinearity among the forecasts that are being combined.

To set notation, suppose that the researcher has available a sample size of T observations and is interested in forecasting the target variable at times $h+1, h+2, ..., T$.

3.1 To Combine or not to Combine

The first question to answer in the presence of multiple forecasts of the same variable is of course whether or not to combine the forecasts or rather simply attempt to identify a single best forecasting model. Here it is important to distinguish between the situation where the information sets underlying the individual forecasts is observed or unobserved to the forecast user. When the information sets are unobserved it is often justified to combine forecasts provided that the private (non-overlapping) parts of the information sets are sufficiently important. Whether this is satisfied can be difficult to assess, but diagnostics such as the correlation between forecasts or forecast errors can be considered.

When forecast users do have access to the full information set used to construct the individual forecasts, Chong and Hendry (1986) and Diebold (1989) argue that combinations may be less justified in the sense that successful combination indicates misspecification of the individual models and so a better individual model should be sought. Finding a 'best' model may of course be rather difficult if the space of models included in the search is high dimensional and the time-series short. As Clemen (1989) nicely puts it: "Using a combination of forecasts amounts to an admission that the forecaster is unable to build a properly specified model. Trying ever more elaborate combining models seems to add insult to injury as the more complicated combinations do not generally perform that well."

Simple tests of whether one forecast dominates another forecast are neither sufficient nor necessary for settling the question of whether or not to combine. This follows since we can construct examples where (in population) forecast $\hat{Y}_1$ dominates forecast $\hat{Y}_2$ (in the sense that it leads to lower expected loss), yet it remains optimal to combine the two forecasts.[4] Similarly, we can construct examples where forecast $\hat{Y}_1$ and $\hat{Y}_2$ generate identical expected loss, yet it is not optimal to combine them−most obviously if they are perfectly correlated, but also due to estimation errors in the combination weights.

What is called for more generally is a test of whether one forecast−or more generally a set of forecasts−encompasses all information contained in another forecast (or sets of forecasts).

[4]Most obviously, under MSE loss, when $Var(Y - \hat{Y}_1) > Var(Y - \hat{Y}_2)$, and $Cor(Y - \hat{Y}_1, Y - \hat{Y}_2) \neq Var(Y - \hat{Y}_2)/Var(Y - \hat{Y}_1)$, it will generally be optimal to combine the two forecasts.

In the context of MSE loss functions, forecast encompassing tests have been developed by Chong and Henry (1986). Point forecasts are sufficient statistics under MSE loss and a test of pair-wise encompassing can be based on the regression

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}_{1t+h,t} + \beta_2 \hat{y}_{2t+h,t} + e_{t+h,t}, \quad t = 1, 2, ...T - h. \tag{31}$$

Forecast 1 encompasses forecast 2 when the parameter restriction $(\beta_0\ \beta_1\ \beta_2) = (0\ 1\ 0)$ holds, while conversely if forecast 2 encompasses forecast 1 we have $(\beta_0\ \beta_1\ \beta_2) = (0\ 0\ 1)$. All other outcomes mean that there is some information in both forecasts which can then be usefully exploited. Notice that this is an argument that only holds in population. It is still possible in small samples that ignoring one forecast can lead to better out-of-sample forecasts even though, asymptotically, the coefficient on the omitted forecast in (31) differs from zero.

More generally, a test that some model, e.g., model 1, forecast encompasses all other forecasts can be based on a test of $\beta_2 = ... = \beta_n$ in the regression

$$y_{t+h} - \hat{y}_{1,t+h,t} = \beta_0 + \sum_{i=2}^{n} \beta_i \hat{y}_{i,t+h,t} + e_{t+h,t}.$$

In situations where the data is not very informative and it is not possible to identify a single dominant model, it makes sense to combine. Makridakis and Winkler (1983) explain this well (page 990): "When a single method is used, the risk of not choosing the best method can be very serious. The risk diminishes rapidly when more methods are considered and their forecasts are averaged. In other words, the choice of the best method or methods becomes less important when averaging." They demonstrate this point by showing that the forecasting performance of a combination strategy improves as a function of the number of models involved in the combination, albeit at a decreasing rate.

Swanson and Teng (2001) propose to use model selection criteria such as the SIC to choose which subset of forecasts to combine. This approach does not require formal hypothesis testing so that size distortions due to the use of sequential test procedures, can be avoided although, of course, consistency of the selection approach must be established for the particular asymptotic experiment appropriate for a given forecasting situation. In empirical work reported by these authors the combination chosen by SIC appears to provide

the best overall performance and rarely gets dominated by other methods in out-of-sample forecasting experiments.

Once it has been established whether to combine or not, there are various ways in which the researcher can estimate the combination weights, $\hat{\omega}_{Th}$. We will discuss some of these methods in turn in what follows. A theme that is common across estimators is that estimation errors in forecast combinations are generally important expecially in cases where N is large relative to T.

3.2 Least Squares Estimators of the Weights

It is common to estimate combination weights by ordinary least squares, regressing realizations of the target variable, Y_τ on the N-vector of forecasts, $\hat{\mathbf{Y}}_\tau$ using data over the period $\tau = h, ..., T$:

$$\hat{\boldsymbol{\omega}}_{Th} = (\sum_{\tau=1}^{T-h} \hat{\mathbf{Y}}_{\tau+h,\tau} \hat{\mathbf{Y}}'_{\tau+h,\tau})^{-1} \sum_{\tau=1}^{T-h} \hat{\mathbf{Y}}_{\tau+h,\tau} Y_{\tau+h}. \tag{32}$$

Many versions of this basic least squares projection have been proposed. Granger and Ramanathan (1984) consider three regressions

$$\begin{aligned} (i)\ Y_{t+h} &= \omega_{0th} + \boldsymbol{\omega}'_{th} \hat{\mathbf{Y}}_{t+h,t} + \varepsilon_{t+h} \\ (ii)\ Y_{t+h} &= \boldsymbol{\omega}'_{th} \hat{\mathbf{Y}}_{t+h,t} + \varepsilon_{t+h} \\ (iii)\ Y_{t+h} &= \boldsymbol{\omega}'_{th} \hat{\mathbf{Y}}_{t+h,t} + \varepsilon_{t+h},\ \text{s.t. } \boldsymbol{\omega}'_{th} \boldsymbol{\iota} = 1. \end{aligned} \tag{33}$$

The first and second of these regressions can be estimated by standard OLS, the only difference being that the second equation omits an intercept term. The third regression omits an intercept and can be estimated through constrained least squares. The first and more general regression does not require that the individual forecasts are unbiased since any bias can be adjusted through the intercept term, ω_{0th}. In contrast, the third specification is motivated by an assumption of unbiasedness of the individual forecasts. Imposing that the weights sum to one then guarantees that the combined forecast is also unbiased. This specification may not be efficient, however, as the latter constraint can lead to efficiency losses as $E[\hat{\mathbf{Y}}_{t+h,t} \varepsilon_{t+h}] \neq \mathbf{0}$. One could further impose convexity constraints $0 \leq \omega_{ith} \leq 1$, $i = 1, .., N$ to rule out that the combined forecast lies outside the range of the individual forecasts.

Another reason for imposing the constraint $\boldsymbol{\omega}_{th}'\boldsymbol{\iota} = 1$ has been discussed by Diebold (1988). He proposes the following decomposition of the forecast error from the combination regression:

$$\begin{aligned} e_{t+h,t}^{c} &= Y_{t+h} - \hat{\omega}_{0th} - \hat{\boldsymbol{\omega}}_{th}'\hat{\mathbf{Y}}_{t+h,t} \\ &= -\hat{\omega}_{0th} + (1 - \hat{\boldsymbol{\omega}}_{th}'\boldsymbol{\iota})Y_{t+h} + \hat{\boldsymbol{\omega}}_{th}'(Y_{t+h}\boldsymbol{\iota} - \hat{\mathbf{Y}}_{t+h,t}) \\ &= -\hat{\omega}_{0th} + (1 - \hat{\boldsymbol{\omega}}_{th}'\boldsymbol{\iota})Y_{t+h} + \hat{\boldsymbol{\omega}}_{th}'\mathbf{e}_{t+h,t}, \end{aligned} \tag{34}$$

where $\mathbf{e}_{t+h,t}$ is the $N \times 1$ vector of h-period forecast errors from the individual models. Oftentimes the target variable, Y_{t+h}, is quite persistent whereas the forecast errors from the individual models are not serially correlated even when $h = 1$. It follows that unless it is imposed that $1 - \hat{\boldsymbol{\omega}}_{th}'\boldsymbol{\iota} = 0$, then the forecast error from the combination regression will typically be serially correlated and hence be predictable itself.

3.3 Recursively Updated Weights

Yang (2004) demonstrates theoretically that linear forecast combinations can lead to far worse performance than those from the single best forecasting model due to large variability in estimates of the combination weights. This happens in part if and when the forecasts are strongly collinear, which is often the case. Yang proposes a range of recursive methods for updating the combination weights that ensure that combinations achieve a performance similar to that of the best individual forecasting method up to a constant penalty term and a proportionality factor. The basic algorithm for calculating combination weights is based on the models' relative forecasting performance:

$$\omega_{i,t+h,t} = \frac{\frac{\pi_i}{\prod_{\tau=1}^{t-h}\hat{\sigma}_{\tau+h,\tau}} \exp\left(\frac{-1}{2}\sum_{\tau=1}^{t-h}\frac{(y_{\tau+h}-\hat{y}_{i,\tau+h,\tau})^2}{\hat{\sigma}_{\tau+h,\tau}}\right)}{\sum_{j=1}^{N}\left(\frac{\pi_j}{\prod_{\tau=1}^{t-h}\hat{\sigma}_{\tau+h,\tau}} \exp\left(\frac{-1}{2}\sum_{\tau=1}^{t-h}\frac{(y_{\tau+h}-\hat{y}_{j,\tau+h,\tau})^2}{\hat{\sigma}_{\tau+h,\tau}}\right)\right)}. \tag{35}$$

Here π_i is the prior weight on model i, while $\hat{y}_{i,\tau+h,\tau}$ is the period-τ forecast from model i for time $\tau + h$ and $\hat{\sigma}_{\tau+h,\tau}^2$ is an estimate of the conditional forecast error variance at time τ which is assumed to be the same across models. This assumption will in general tend to favor combination schemes with less dispersion across the weights. Yang derives

bounds on the MSE of the combined forecasts, $\sum_{j=1}^{N} \omega_{j,\tau+h,\tau}\hat{y}_{j,\tau+h,\tau}$ relative to the MSE of the best individual model included in the set of forecasts considered in the combination. Let $m_{t+h,t}$ be the true but unknown conditional mean of Y_{t+h} given some information set $\mathcal{I}_t$ and assume that the combined forecast is formed as a convex combination of a set of forecasting procedures all adapted to $\mathcal{I}_t$ using the weights in (35),

$$\hat{y}^c_{t+h,t} = \sum_{j=1}^{N} \omega_{j,t+h,t}\hat{y}_{j,t+h,t}. \tag{36}$$

Under the assumption that the following condition holds

$$\sup_{j\geq 1} \frac{|\hat{y}_{j,t+h,t} - m_{t+h,t}|}{\sigma_{t+h,t}} \leq \sqrt{\phi},$$

Yang shows that−under normality of the forecast errors and assuming that $\sigma_{t+h,t}$ is known−the mean average square risk (relative to the conditional variance) of the combination (36) satisfies

$$\frac{1}{T}\sum_{\tau=1}^{T} E\left[\frac{(\hat{y}^c_{\tau+h,\tau} - m_{\tau+h,\tau})^2}{\sigma^2_{\tau+h,t}}\right] \leq \left(2 + \frac{9\phi}{2}\right) \inf_{j\geq 1}\left(\frac{2\log(1/\pi_j)}{T} + \frac{1}{T}\sum_{\tau=1}^{T} E\left[\frac{(\hat{y}_{j,\tau+h,\tau} - m_{\tau+h,\tau})^2}{\sigma^2_{\tau+h,\tau}}\right]\right).$$

Hence the combined forecast under this updating method achieves the same performance as the optimal model, M_j, $j = 1, ..., N$ up to a constant factor and an additive penalty term, $\log(1/\pi_j)/T$. The result simplifies when T rises in which case the individual forecasting models get closer to the conditional means, $m_{\tau+h,\tau}$. This allows Yang to establish that

$$\limsup_{T\to\infty} \left\{ \frac{\frac{1}{T}\sum_{\tau=1}^{T} E\left[\frac{(\hat{y}^c_{\tau+h,\tau} - m_{\tau+h,\tau})^2}{v_{\tau+h,\tau}}\right]}{\inf_{j\geq 1}\left(\frac{1}{T}\sum_{\tau=1}^{T} E\left[\frac{(\hat{y}_{j,\tau+h,\tau} - m_{\tau+h,\tau})^2}{v_{\tau+h,\tau}}\right]\right)} \right\} \leq 2.$$

This means that, in large samples, the performance of the combined forecast (36) will be at least half as good as the performance of the best individual model. Of course, this need not be the best achievable result−it is possible that better combinations could be found that account for the correlation structure across forecast errors which are ignored in the combination scheme (36).

Yang (2004) also characterizes the cost of linear combination theoretically. The risk of the combination, $R(\hat{y}^c_{T+1,T}; T)$ relative to the risk of the best individual model (whose identity

is unknown), $R(M^*;T)$, is bounded by

$$R(\hat{y}^c_{T+1,T};T) \leq C \begin{cases} R(M^*;T) + \frac{N\log(1+T/N)}{T^{1-\tau}} & \text{when } 1 \leq N < \sqrt{T} \\ R(M^*;T) + \frac{\log(N)}{\sqrt{T\log(T)}} & \text{when } N \geq \sqrt{T} \end{cases},$$

where C is a constant depending on technical assumptions outlined in Yang (2004). The associated 'complexity' or penalty term

$$\psi_T(N) = \begin{cases} \frac{N\log(1+T/N)}{T^{1-\tau}} & \text{when } 1 \leq N < \sqrt{T} \\ \frac{\log(N)}{\sqrt{T\log(T)}} & \text{when } N \geq \sqrt{T} \end{cases}$$

is the cost of not knowing the correct combination weights in the linear combination (36). It is an increasing function of the number of models, N, and a decreasing function of T. When T is small the cost of combining many methods (i.e. keeping N large) can be very substantial.

These results are extended to the case where $\sigma^2_{t+h,t}$ has to be estimated and to the case where errors are non-Gaussian, which requires kernel estimation. Since constructing the weights (35) requires estimating the conditional variances $\sigma_{t+h,t}$, Yang proposes an alternative approach that does not require such estimates and is instead based on the loss $L(y_{\tau+h,\tau} - \hat{y}_{i,\tau+h,\tau})$:

$$\omega_{i,\tau+h,\tau} = \frac{\pi_i \exp\left(-\lambda \sum_{\tau=1}^{t-h} L(y_{\tau+h,\tau} - \hat{y}_{i,\tau+h,\tau})\right)}{\sum_j \left(\pi_j \exp\left(-\lambda \sum_{\tau=1}^{t-h} L(y_{\tau+h,\tau} - \hat{y}_{j,\tau+h,\tau})\right)\right)}.$$

3.4 Relative Performance Weights

Estimation errors in the combination weights tend to be particularly large due to difficulties in precisely estimating the full covariance matrix, $\mathbf{\Sigma}_e$. One answer to this problem is to simply ignore correlations across forecast errors. Combination weights that reflect the performance of each individual model relative to the performance of the average model, but ignore correlations across forecasts have been proposed by Bates and Granger (1969) and Newbold and Granger (1974). Both papers argue that correlations can be poorly estimated and should be ignored in situations with many forecasts and short time-series. This effectively amounts to treating $\mathbf{\Sigma}_e$ as a diagonal matrix, c.f. Winkler and Makridakis (1983).

Stock and Watson (2001) propose a broader set of combination weights that also ignore correlations between forecast errors but base the combination weights on the models' relative MSE performance raised to various powers. Let $MSE_{i,t+h,t} = (1/v)\sum_{\tau=t-v}^{t} e_{i,\tau,\tau-h}^2$ be the ith forecasting model's MSE at time t, computed over a window of the previous v periods. Then

$$\begin{aligned} \hat{y}_{t+h,t}^{c} &= \sum_{i=1}^{N} \omega_{i,t+h,t}\hat{y}_{i,t+h,t} \\ \omega_{i,t+h,t} &= \frac{(1/MSE_{i,t+h,t}^{\kappa})}{\sum_{j=1}^{N}(1/MSE_{j,t+h,t}^{\kappa})}. \end{aligned} \tag{37}$$

Setting $\kappa = 0$ assigns equal weights to all forecasts, while forecasts are weighted by the inverse of their MSE when $\kappa = 1$. The latter strategy has been found to work well in practice as it does not require estimating the off-diagonal parameters of the covariance matrix of the forecast errors. Such weights therefore disregard any correlations between forecast errors and so are only optimal in large samples provided that the forecast errors are truly uncorrelated.

3.5 Moment Estimators

Outside the quadratic loss framework one can base estimation of the combination weights directly on the loss function, c.f. Elliott and Timmermann (2004). Let the realized loss in period $t+h$ be $L(e_{t+h}; \boldsymbol{\omega}, \boldsymbol{\theta}_L)$

$$L(e_{t+h}; \boldsymbol{\omega}) = L(\boldsymbol{\omega} | y_{t+h}, \widehat{\mathbf{y}}_{t+h,t}, \boldsymbol{\theta}_L),$$

where $\boldsymbol{\theta}_L$ are the parameters of the loss function. Then ω_0 and $\boldsymbol{\omega}$ can be obtained as an M-estimator based on the sample analog of $E[L(e_{t+h})]$ using a sample of $T-h$ observations $\{y_\tau, \widehat{\mathbf{y}}_{\tau,\tau-h}\}_{\tau=h+1}^{T}$:

$$\bar{L}(\boldsymbol{\omega}) = T^{-1}\sum_{\tau=h+1}^{T} L(e_{\tau,\tau-h}(\boldsymbol{\omega}); \boldsymbol{\theta}_L).$$

Taking derivatives, one can use the generalized method of moments (GMM) to estimate $\boldsymbol{\omega}_{Th}$ from the quadratic form

$$\min_{\boldsymbol{\omega}_{Th}} \left(\sum_{\tau=h+1}^{T} \mathbf{L}'(e_{\tau,\tau-h}(\boldsymbol{\omega}_{Th}); \boldsymbol{\theta}_L)\right)' \boldsymbol{\Sigma}^{-1} \left(\sum_{\tau=h+1}^{T} \mathbf{L}'(e_{\tau,\tau-h}(\boldsymbol{\omega}_{Th}); \boldsymbol{\theta}_L)\right), \tag{38}$$

where $\mathbf{\Sigma}$ is a (positive definite) weighting matrix. Consistency and asymptotic normality of the estimated weights is easily established under standard regularity conditions.

3.6 Non-parametric Combination Schemes

The estimators considered so far require stationarity at least for the moments involved in the estimation. To be successful, they also require a reasonably large data sample (relative to the number of models, N) as they otherwise tend not to be robust to outliers, c.f. Gupta and Wilton (1987) p. 358: "...combination weights derived using minimum variance or regression are not robust given short data samples, instability or nonstationarity. This leads to poor performance in the prediction sample." In many applications the number of forecasts, N, is large relatively to the length of the time-series, T. In this case, it is not feasible to estimate the combination weights by OLS. Simple combination schemes such as an equal-weighted average of forecasts $y^{ew}_{t+h,t} = \boldsymbol{\iota}'\hat{\mathbf{y}}_{t+h,t}/N$ or weights based on the inverse MSE-values often are an attractive option in this situation.

Simple, rank-based weighting schemes can also be constructed and have been used with some success in mean-variance analysis in finance, c.f. Wright and Satchell (2003). These take the form $\boldsymbol{\omega}_{th} = f(\mathcal{R}_{1,t,t-h}, ..., \mathcal{R}_{N,t,t-h})$, where $\mathcal{R}_{i,t,t-h}$ is the rank of the ith model based on its $h-$period performance up to time t. The most common scheme in this class is to simply use the median forecast as proposed by authors such as Armstrong (1989), Hendry and Clements (2002) and Stock and Watson (2001, 2003).

The triangular weighting (TK) scheme lets the combination weights to be inversely proportional to the models' rank:

$$\omega_{i,t+h,t} = \mathcal{R}^{-1}_{i,t+h,t}/(\sum_{i=1}^{N} \mathcal{R}^{-1}_{i,t+h,t}). \tag{39}$$

Again this combination ignores correlations across forecast errors. However, since ranks are likely to be less sensitive to outliers, this weighting scheme can be expected to be more robust than the weights in (37).

Another example in this class is spread combinations. These have been proposed by

Aiolfi and Timmermann (2004) and assume weights of the form

$$\omega_{i,t+h,t} = \begin{cases} \frac{1+\bar{\omega}}{\alpha N} & \text{if } \mathcal{R}_{i,t+h,t} \leq \alpha N \\ 0 & \text{if } \alpha N < \mathcal{R}_{i,t+h,t} < (1-\alpha)N \\ \frac{-\bar{\omega}}{\alpha N} & \mathcal{R}_{i,t+h,t} \leq (1-\alpha)N \end{cases}, \tag{40}$$

where α is the proportion of top models that - based on performance up to time t - gets a weight of $(1+\bar{\omega})/\alpha N$. Similarly, a proportion α of models gets a weight of $-\bar{\omega}/\alpha N$. The larger the value of α, the wider the set of top and bottom models that are used in the combination. Similarly, the larger is $\bar{\omega}$, the bigger the difference in weights on top and bottom models. Such spreads are likely to work well if the correlation between forecast errors grows and the larger the forecast error is. The intuition for such spread combinations can be seen from (12) when $N = 2$ so $\alpha = 1/2$. Solving for ρ_{12} we see that $\omega^* = 1 + \bar{\omega}$ provided that

$$\rho_{12} = \frac{1}{2\bar{\omega}+1}\left(\frac{\sigma_2}{\sigma_1}\bar{\omega} + \frac{\sigma_1}{\sigma_2}(1+\bar{\omega})\right).$$

Hence if $\sigma_1 \approx \sigma_2$, spread combinations are close to optimal provided that $\rho_{12} \approx 1$. The second forecast provides a hedge for the performance of the first forecast in this situation. In general, spread portfolios are likely to work well when the forecasts are strongly multicollinear.

Bunn (1975) proposes an outperformance scheme that is based on the probability that a model performs best in the preceding sample, i.e.

$$\begin{aligned} p_{it+h,t} &= \Pr(L(e_{it+h,t}) < L(e_{jt+h,t})) \text{ for all } j \neq i \\ \hat{y}^c_{t+h,t} &= \sum_{i=1}^{N} p_{it+h,t}\hat{y}_{it+h,t}. \end{aligned}$$

Bunn discusses how $p_{it+h,t}$ can be updated based on a model's track historical record using the proportion of times up to the current period where a model outperformed its competitors.

Gupta and Wilson (1987) propose an odds matrix approach based on a matrix of pairwise odds on outperformance to derive combination weights. Let π_{ij} be the probability that the ith forecasting model outperforms the jth model next period. The ratio $o_{ij} = \pi_{ij}/\pi_{ji}$ is then the odds that model i will outperform model j and $o_{ij} = 1/o_{ji}$. Filling out the $N \times N$ odds ratio matrix O with i, j element o_{ij} requires specifying $N(N-1)/2$ pairs of probabilities

of outperformance, π_{ij}. An estimate of the combination weight ω is obtained by solving a system of N linear equations, $(O - kI)\omega = \mathbf{0}$. Since O has unit rank, ω can be found as the normalized eigenvector associated with the largest (and only non-zero) eigenvalue of O. This approach gives weights that are insensitive to small changes in the odds ratio and so does not require large amounts of data. Also, as it does not account for dependencies between the models it is likely to be less sensitive to changes in the covariance matrix than the regression approach. Conversely, it will perform worse if such correlations are important and can be estimated with sufficient precision.

3.7 Pooling, Clustering and Trimming

Rather than combining the full set of forecasts, it is often advantageous to discard the models with the worst performance (trimming). Combining only the best models goes under the header 'use sensible models' in Armstrong (1989). This is particularly important when forecasting with nonlinear models whose predictions are often implausible and lie outside the empirical range of the target variable. One can base whether or not to trim−and by how much to trim−on formal tests or on more lose decision rules.

To see why trimming can be important, suppose a fraction α of the forecasting models contain valuable information about the target variable while a fraction $1-\alpha$ is pure noise. It is easy to see in this extreme case that the optimal forecast combination sets zero weight on the pure noise forecasts. However, once combination weights have to be estimated, forecasts that only add marginal information should better be dropped from the combination since the cost of their inclusion−increased parameter estimation error−does not match the benefit.

The 'thick modeling' approach−thus named because it seeks to exploit information in a cross-section (thick set) of models−proposed by Granger and Jeon (2004) is an example of a trimming scheme that removes poorly performing models in a step that precedes calculation of combination weights. Granger and Jeon argue that "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."

Grouping or clustering of forecasts can be motivated by the assumption of a common

factor structure underlying the forecasting models. Consider the factor model

$$\begin{aligned} Y_{t+h} &= \mu_y + \boldsymbol{\beta}_y' \mathbf{F}_{t+h} + \varepsilon_{yt+h}, \\ \hat{\mathbf{Y}}_{t+h,t} &= \boldsymbol{\mu} + \mathbf{B}\mathbf{F}_{t+h} + \boldsymbol{\varepsilon}_{t+h}, \end{aligned} \tag{41}$$

where $\mathbf{F}_{t+h}$ is an $n_f \times 1$ vector of factor realizations satisfying $E[\mathbf{F}_{t+h}\varepsilon_{yt+h}] = \mathbf{0}$, $E[\mathbf{F}_{t+h}\boldsymbol{\varepsilon}_{t+h}'] = \mathbf{0}$ and $E[\mathbf{F}_{t+h}\mathbf{F}_{t+h}'] = \boldsymbol{\Sigma}_F$. $\boldsymbol{\beta}_y$ is an $n_f \times 1$ vector while $\mathbf{B}$ is an $N \times n_f$ matrix of factor loadings. For simplicity we can assume that the factors have been orthogonalized. This will obviosly hold if they are constructed as the principal components from a large data set and can otherwise be achieved through rotation. Furthermore, all innovations ε are serially uncorrelated with zero mean, $E[\varepsilon_{yt+h}^2] = \sigma_{\varepsilon_y}^2$, $E[\varepsilon_{yt+h}\boldsymbol{\varepsilon}_{t+h}] = \mathbf{0}$ and the 'noise' in the forecasts is idiosyncratic (model specific), i.e.,

$$E[\varepsilon_{it+h}\varepsilon_{jt+h}] = \begin{cases} \sigma_{\varepsilon_i}^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

We arrange these values on a diagonal matrix $E[\boldsymbol{\varepsilon}_{t+h}\boldsymbol{\varepsilon}_{t+h}'] = \mathbf{D}_\varepsilon$. This gives the following moments

$$\begin{pmatrix} Y_{t+h} \\ \hat{\mathbf{Y}}_{t+h,t} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\beta}_y'\boldsymbol{\Sigma}_F\boldsymbol{\beta}_y + \sigma_{\varepsilon_Y}^2 & \boldsymbol{\beta}_y'\boldsymbol{\Sigma}_F\mathbf{B}' \\ \mathbf{B}\boldsymbol{\Sigma}_F\boldsymbol{\beta}_y & \mathbf{B}\boldsymbol{\Sigma}_F\mathbf{B}'+\mathbf{D}_\varepsilon \end{pmatrix} \right).$$

Also suppose either that $\boldsymbol{\mu} = \mathbf{0}$, $\mu_y = 0$ or a constant is included in the combination scheme. Then the first order condition for the optimal weights is, c.f. (8),

$$\boldsymbol{\omega}^* = (\mathbf{B}\boldsymbol{\Sigma}_F\mathbf{B}'+\mathbf{D}_\varepsilon)^{-1}\mathbf{B}\boldsymbol{\Sigma}_F\boldsymbol{\beta}_y. \tag{42}$$

Further suppose that the N forecasts of the n_f factors can be divided into appropriate groups according to their factor loading vectors $\mathbf{b}_i$ such that $\sum_{i=1}^{n_f} \dim(\mathbf{b}_i) = N$

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_2 & \mathbf{0} & \cdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{b}_{n_f} \end{pmatrix}.$$

Then

$$\mathbf{B}\boldsymbol{\Sigma}_F\mathbf{B}'+\mathbf{D}_{\boldsymbol{\varepsilon}} = \begin{pmatrix} \mathbf{b}_1\mathbf{b}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_2\mathbf{b}_2' & \mathbf{0} & \cdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{b}_{n_f}\mathbf{b}_{n_f}' \end{pmatrix} \mathbf{D}_{\sigma_F^2}+\mathbf{D}_{\boldsymbol{\varepsilon}}, \tag{43}$$

where $\mathbf{D}_{\sigma_F}$ is a diagonal matrix with $\sigma^2_{F_1}$ in its first n_1 diagonal places followed by $\sigma^2_{F_2}$ in the next n_2 diagonal places and so on and $\mathbf{D}_{\varepsilon}$ is a diagonal matrix with $Var(\varepsilon_{it})$ as the ith diagonal element. Thus the matrix (43) will be block diagonal and hence its inverse will also be block diagonal. Provided that the forecasts tracking the individual factors can be grouped and have similar factor exposure ($\mathbf{b}_i$) within each group, this suggests that little is lost by pooling forecasts within each cluster and ignoring correlations across clusters. In a subsequent step, sample counterparts of the optimal combination weights for the grouped forecasts can be obtained by least-squares estimation. In this way, far fewer combination weights (n_f rather than N) have to be obtained. This is expected to decrease forecast errors and thus improve forecasting performance.

Building on these ideas Aiolfi and Timmermann (2004) propose to sort forecasting models into clusters based on their past MSE performance but, as the previous argument suggests, one could alternatively base clustering on correlation patterns among the forecast errors.[5] Their method identifies K clusters. Let $\hat{\mathbf{y}}^k_{t+h,t}$ be the $p_k \times 1$ vector containing the subset of forecasts belonging to cluster k, $k = 1, 2, .., K$. By ordering the clusters such that the first cluster contains models with the lowest historical MSFE values, Aiolfi and Timmermann consider three separate strategies. The first simply computes the average forecast across models in the cluster of previous best models:

$$\hat{y}^{CPB}_{t+h,t} = (\boldsymbol{\iota}'_{p_1}/p_1)\hat{\mathbf{y}}^1_{t+h,t} \tag{44}$$

The second combination strategy identifies a small number of clusters, pools forecasts within each cluster and then estimates optimal weights on these pooled predictions by least

[5] The two clustering methods will be similar if σ_{F_i} varies significantly across factors and the factor exposure vectors, $\mathbf{b}_i$, and error variances $\sigma^2_{\varepsilon_i}$ are not too dissimilar across models. In this case forecast error variances will tend to cluster around the factors that the various forecasting models are most exposed to.

squares:

$$\hat{y}_{t+h,t}^{CLS} = \sum_{k=1}^{K} \hat{\omega}_{kt} \left[(\boldsymbol{\iota}'_{p_k}/p_k) \hat{\mathbf{y}}_{t+h,t}^{k} \right], \tag{45}$$

where $\hat{\omega}_{kt}$ are least-squares estimates of the optimal combination weights for the K clusters. This strategy is likely to work well if the variation in forecasting performance within each cluster is small relative to the variation in forecasting performance across clusters.

Finally, the third strategy pools forecasts within each cluster, estimates least squares combination weights and then shrinks these towards equal weights in order to reduce the effect of parameter estimation error

$$\hat{y}_{t+h,t}^{CSW} = \sum_{k=1}^{K} \hat{s}_{kt} \left[(\boldsymbol{\iota}'_{p_k}/p_k) \hat{\mathbf{y}}_{t+h,t}^{k} \right],$$

where $\hat{s}_{kt}$ are the shrinkage weights for the K clusters computed as: $\hat{s}_{kt} = \lambda \hat{\omega}_{kt} + (1-\lambda)\frac{1}{K}$, $\lambda = \max\left\{0, 1 - \kappa\left(\frac{K}{t-h-K}\right)\right\}$. The higher is κ, the higher the shrinkage towards equal weights.

4 Time-varying and Nonlinear combination Methods

Two families of combination schemes of special interest that generalize (6) are linear combinations with time-varying weights:

$$\hat{y}_{t+h,t}^{c} = \omega_{0th} + \boldsymbol{\omega}'_{th} \widehat{\mathbf{y}}_{t+h,t}, \tag{46}$$

where ω_{0th}, $\boldsymbol{\omega}'_{th}$ are adapted to $\mathcal{I}_t$ and non-linear combinations with constant weights:

$$\hat{y}_{t+h,t}^{c} = g(\widehat{\mathbf{y}}_{t+h,t}, \boldsymbol{\theta}), \tag{47}$$

where $g(.)$ is some function that is nonlinear in the parameters, $\boldsymbol{\theta}$, in the vector of forecasts, $\widehat{\mathbf{y}}_{t+h,t}$, or in both. Naturally, there is a close relationship between time-varying and nonlinear combinations. For example, non-linearities in the true data generating process can lead to time-varying covariances for the forecast errors and hence time-varying weights in the combination of (misspecified) forecasts.

We next describe some of the approaches within these classes that have been proposed in the literature.

4.1 Time-varying Weights

When the joint distribution of $(Y_{t+h}\ \hat{\mathbf{Y}}'_{t+h,t})'$, or at least its first and second moments, vary over time, it can be beneficial to let the combination weights change over time. Indeed, Bates and Granger (1969) suggested either assigning a disproportionately large weight to the model that has performed best most recently or using an adaptive updating scheme that puts more emphasis on recent performance in assigning the combination weights. Rather than explicitly modeling the structure of the time-variation in the combination weights, Bates and Granger proposed five adaptive estimation schemes based on exponential discounting or the use of rolling estimation windows.

The first combination scheme, later generalized in Newbold and Granger (1974), uses a rolling window of the most recent v observations based on the forecasting models' relative performance

$$\omega_{it}^{NG1} = \frac{\left(\sum_{\tau=t-v+1}^{t} e_{i\tau}^2\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-v+1}^{t} e_{j\tau}^2\right)^{-1}}. \tag{48}$$

The shorter is v, the more weight is put on the models' recent track record and the larger the part of the historical data that is discarded. If $v = t$, an expanding window is used and this becomes a special case of (37). Clearly, correlations between forecast errors are ignored by this scheme.

The second rolling window scheme accounts for correlations across forecast errors but, again, only uses the most recent v observations for estimation:

$$\begin{aligned} \boldsymbol{\omega}_t^{NG2} &= \hat{\boldsymbol{\Sigma}}_{et}^{-1}\boldsymbol{\iota}/(\boldsymbol{\iota}'\hat{\boldsymbol{\Sigma}}_{et}^{-1}\boldsymbol{\iota}), \\ \hat{\boldsymbol{\Sigma}}_{et}[i,j] &= v^{-1}\sum_{\tau=t-v}^{t-1} e_{i\tau}e_{j\tau}. \end{aligned} \tag{49}$$

The third combination scheme uses adaptive updating captured by the parameter $\alpha \in (0;1)$, which tends to smooth the time-series evolution in the combination weights:

$$\omega_{it}^{NG3} = \alpha\omega_{it-1} + (1-\alpha)\frac{\left(\sum_{\tau=t-v}^{t-1} e_{i\tau}^2\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-v}^{t-1} e_{j\tau}^2\right)^{-1}}. \tag{50}$$

The closer to unity is α, the smoother the weights will generally be.

The fourth and fifth combination methods are based on exponential discounting versions of the first two methods and take the form

$$\omega_{it} = \frac{\left(\sum_{\tau=1}^{t-1} \lambda^{\tau} e_{i\tau}^2\right)^{-1}}{\sum_{j=1}^{N} \left(\sum_{\tau=1}^{t-1} \lambda^{\tau} e_{j\tau}^2\right)^{-1}}, \tag{51}$$

where $\lambda \geq 1$ and higher values of λ correspond to putting more weight on recent data. This scheme does not put a zero weight on any of the past forecast errors whereas the rolling window methods entirely ignore observations more than v periods old. If $\lambda = 1$, there is no discounting of past performance and the formula becomes a special case of (37). However, it is common to use a discount factor such as $\lambda = 0.95$ or $\lambda = 0.90$, although the chosen value will depend on factors such as data frequency, evidence of instability etc.

Finally, the fifth scheme estimates the variance and covariance of the forecast errors using exponential discounting:

$$\begin{aligned} \boldsymbol{\omega}_t &= \hat{\boldsymbol{\Sigma}}_{et}^{-1} \boldsymbol{\iota} / (\boldsymbol{\iota}' \hat{\boldsymbol{\Sigma}}_{et}^{-1} \boldsymbol{\iota}), \\ \hat{\boldsymbol{\Sigma}}_{et}[i,j] &= \sum_{\tau=1}^{t-1} \lambda^{\tau} e_{i\tau} e_{j\tau}. \end{aligned} \tag{52}$$

Putting more weight on recent data means reducing the weight on past data and tends to increase the variance of the parameter estimates. Hence it will typically lead to poorer performance if the underlying data generating process is truly covariance stationary. Conversely, the underlying time-variations have to be quite strong to justify not using an expanding window.

Diebold and Pauly (1987) embed these schemes in a general weighted least squares setup that chooses combination weights to minimize the weighted average of forecast errors

$$\sum_{t=1}^{T} \sum_{\tau=1}^{T} \omega_{t,\tau} e_t e_{\tau}, \tag{53}$$

or equivalently, $\mathbf{e}'\mathbf{W}\mathbf{e}$, where $\mathbf{W}$ is a $T \times T$ matrix with $[t, \tau]$ element $\omega_{t,\tau}$. Assuming that $\mathbf{W}$ is diagonal, equal-weights correspond to $\omega_{tt} = 1$ for all t, linearly declining weights can be represented as $\omega_{tt} = t$, and geometrically declining weights take the form $\omega_{tt} = \lambda^{T-t}$, $0 < \lambda \leq 1$. Finally, they introduce two new weighting schemes, namely nonlinearly declining

weights, $\omega_{tt} = t^{\lambda}$, $\lambda \geq 0$ and the Box-Cox transform weights

$$\omega_{tt} = \begin{cases} (t^{\lambda} - 1)/\lambda & \text{if } 0 < \lambda \leq 1 \\ \ln(t) & \text{if } \lambda = 0 \end{cases}.$$

The weights $\omega_{tt} = t^{\lambda}$ can be either declining at an increasing rate or at a decreasing rate, depending on the sign of $\lambda - 1$. This is clearly an attractive feature and one that, e.g., the geometrically declining weights do not have.

Diebold and Pauly also consider regerssion-based combinations with time-varying parameters. For example, if the time-variation in the combination weights can be modeled through a deterministic polynomial, we get

$$y_{t+h} = \sum_{i=0}^{r} (\boldsymbol{\beta}_i' t^i) \hat{\mathbf{y}}_{t+h,t} + \varepsilon_{t+h},$$

where $\boldsymbol{\beta}_i'$ is an $m \times 1$ vector of combination weights on the ith polynomial term. This approach explicitly models the evolution in the combination weights as opposed to doing this indirectly through the weighting of past and current forecast errors.

Instead of using adaptive schemes for updating the parameter estimates, an alternative is to explicitly model time-variations in the combination weights. One class of combination schemes considered by, e.g., Sessions and Chatterjee (1989), Zellner, Hong and Min (1991) and Lesage and Magura (1992) lets the combination weights evolve according to a time-varying parameter model:

$$\begin{aligned} y_{t+h} &= \widetilde{\boldsymbol{\omega}}_{t+h,t}' \mathbf{z}_{t+h} + \varepsilon_{t+h}, \\ \widetilde{\boldsymbol{\omega}}_{t+h,t} &= \widetilde{\boldsymbol{\omega}}_{t,t-h} + \boldsymbol{\eta}_{t+h}, \end{aligned} \tag{54}$$

where $\mathbf{z}_{t+h} = (1 \ \widehat{\mathbf{y}}_{t+h,t}')'$ and $\widetilde{\boldsymbol{\omega}}_{t+h,t} = (\omega_{0t+h,t} \ \boldsymbol{\omega}_{t+h,t})$. It is typically assumed that $\varepsilon_{t+h} \sim iid(0, \sigma_{\varepsilon}^2)$, $\boldsymbol{\eta}_{t+h} \sim iid(0, \boldsymbol{\sigma}_{\eta}^2)$ and $Cov(\varepsilon_{t+h}, \boldsymbol{\eta}_t) = \mathbf{0}$.

Changes in the combination weights may instead occur more discretely, driven by some switching indicator, $I_{\mathbf{e}}$, c.f. Deutsch, Granger and Terasvirta (1994):

$$y_{t+h} = I_{\mathbf{e}_t \in A}(\omega_{01} + \boldsymbol{\omega}_1' \widehat{\mathbf{y}}_{t+h,t}) + (1 - I_{\mathbf{e}_t \in A})(\omega_{02} + \boldsymbol{\omega}_2' \widehat{\mathbf{y}}_{t+h,t}) + \varepsilon_{t+h}. \tag{55}$$

Here $\mathbf{e}_t = \boldsymbol{\iota} y_t - \widehat{\mathbf{y}}_{t,t-h}$ is the vector of period-t forecast errors; $I_{\mathbf{e}_t \in A}$ is an indicator function taking the value unity when $\mathbf{e}_t \in A$ and zero otherwise, for A some pre-defined set defining

the switching condition. This provides a broad class of time-varying combination schemes as $I_{\mathbf{e}_t \in A}$ can depend on past forecast errors or other variables in a number of ways. For example, $I_{\mathbf{e}_t \in A}$ could be unity if the forecast error is positive, zero otherwise.

Engle, Granger and Kraft (1984) propose time-varying combining weights that follow a bivariate ARCH scheme. They assume that the distribution of the two forecast errors $\mathbf{e}_{t+h,t} = (e_{1t+h,t} \; e_{2t+h,t})'$ is bivariate Gaussian $N(\mathbf{0}, \mathbf{H}(\mathcal{I}_t))$ where $\mathbf{H}(\mathcal{I}_t)$ has i,j element H_{ijt} and depends on powers of past forecast errors. The combination scheme considered by Engle, Granger and Kraft constrains the weights to sum to unity so the forecast error becomes $e_{t+h} = \omega_{t+h,t} e_{1,t+h,t} + (1 - \omega_{t+h,t}) e_{2,t+h,t}$, where

$$\omega_{t+h,t}^{EGK} = \frac{H_{22t} - H_{12t}}{H_{11t} + H_{22t} - 2H_{12t}},$$

where, in the basic combination scheme with eight lags,

$$\begin{pmatrix} H_{11t} \\ H_{21t} \\ H_{22t} \end{pmatrix} = \begin{pmatrix} a_{01} \\ a_{02} \\ a_{03} \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{8} e_{1t-j,t-j-h}^2 \\ \sum_{j=1}^{8} e_{1t-j,t-j-h} e_{2t-j,t-j-h} \\ \sum_{j=1}^{8} e_{1t-j,t-j-h}^2 \end{pmatrix}, \tag{56}$$

which is the direct generalization of (12). In their empirical application to inflation forecasting, the associated combination weights turn out to be quite unstable and dominated by estimation error so a second scheme that includes both the average forecast error, $(e_{1,t+h,t} + e_{2,t+h,t})/2$ and squared deviations, $(e_{1,t+1,t} - e_{2,t+1,t})^2$ is proposed:

$$\begin{aligned} H_{11t} &= a_{01} + a_{11} \sum_{j=1}^{8} (e_{1,t-j,t-j-h} + e_{2,t-j,t-j-h})^2 + a_{12} \sum_{j=1}^{8} (e_{1,t-j,t-j-h} - e_{2,t-j,t-j-h})^2 \\ H_{11t} &= a_{02} + a_{21} \sum_{j=1}^{8} (e_{1,t-j,t-j-h} + e_{2,t-j,t-j-h})^2 \\ H_{11t} &= a_{03} + a_{31} \sum_{j=1}^{8} (e_{1,t-j,t-j-h} + e_{2,t-j,t-j-h})^2 + a_{32} \sum_{j=1}^{8} (e_{1,t-j,t-j-h} - e_{2,t-j,t-j-h})^2. \end{aligned} \tag{57}$$

This leads to smoother combination weights and better out-of-sample forecasting performance.[6]

[6] In their application Engle, Granger and Kraft set $h = 1$ and thus consider combinations of one-step-ahead forecasts.

A final model for time-variation in the combination weights has been proposed by Elliott and Timmermann (2003). This approach is able to track both sudden and discrete as well as more gradual shifts in the joint distribution of $(Y_{t+h} \ \hat{\mathbf{Y}}'_{t+h,t})'$. Suppose that the joint distribution of $(Y_{t+h} \ \hat{\mathbf{Y}}'_{t+h,t})$ is driven by an unobserved state variable, S_{t+h}, which assumes one of n_s possible values, i.e. $S_{t+h} \in (1, ..., n_s)$. Conditional on a given realization of the underlying state, $S_{t+h} = s_{t+h}$, the joint distribution of Y_{t+h} and $\hat{\mathbf{Y}}_{t+h}$ is assumed to be Gaussian

$$\left(\begin{array}{c} Y_{t+h} \\ \hat{\mathbf{Y}}_{t+h,t} \end{array} \right) \Bigg|_{s_{t+h}} \sim N \left(\left(\begin{array}{c} \mu_{y s_{t+h}} \\ \boldsymbol{\mu}_{\hat{\mathbf{y}} s_{t+h}} \end{array} \right), \left(\begin{array}{cc} \sigma^2_{y s_{t+h}} & \boldsymbol{\sigma}'_{y \hat{\mathbf{y}} s_{t+h}} \\ \boldsymbol{\sigma}_{y \hat{\mathbf{y}} s_{t+h}} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}} \hat{\mathbf{y}} s_{t+h}} \end{array} \right) \right), \tag{58}$$

which is similar to (7) but now conditional on S_{t+h}, which is important. It generalizes (28) to allow for an arbitrary number of states. State transitions are assumed to be driven by a first-order Markov chain $\mathbf{P} = \Pr(S_{t+h} = s_{t+h} | S_t = s_t)$

$$\mathbf{P} = \left(\begin{array}{cccc} p_{11} & p_{12} & \cdots & p_{1 n_s} \\ p_{21} & p_{22} & \cdots & \vdots \\ \vdots & \vdots & \cdots & p_{n_s - 1 n_s} \\ p_{n_s 1} & \cdots & p_{n_s n_s - 1} & p_{n_s n_s} \end{array} \right). \tag{59}$$

Conditional on $S_{t+h} = s_{t+h}$, the expectation of Y_{t+h} is linear in the prediction signals, $\hat{\mathbf{Y}}_{t+h,t}$, and thus takes the form of state-dependent intercept and combination weights:

$$E[Y_{t+h} | \hat{\mathbf{y}}_{t+h,t}, s_{t+h}] = \mu_{y s_{t+h}} + \boldsymbol{\sigma}'_{y \hat{\mathbf{y}} s_{t+h}} \boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}} \hat{\mathbf{y}} s_{t+h}} (\hat{\mathbf{y}}_{t+h,t} - \boldsymbol{\mu}_{\hat{\mathbf{y}} s_{t+h}}). \tag{60}$$

Accounting for the fact that the underlying state is unobservable, the conditionally expected loss given current information, $\mathcal{I}_t$, becomes:

$$E\left[e^2_{t+h} | \mathcal{I}_t\right] = \sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t} \left\{ \mu^2_{e s_{t+h}} + \sigma^2_{e s_{t+h}} \right\}, \tag{61}$$

where $\pi_{s_{t+h},t} = \Pr(S_{t+h} = s_{t+h} | \mathcal{I}_t)$ is the probability of being in state s_{t+h} in period $t+h$ conditional on current information, $\mathcal{I}_t$. Assuming a linear combination conditional on $\mathcal{I}_t$,

the optimal combination weights, $\omega^*_{0th}, \boldsymbol{\omega}^*_{th}$ become (c.f. Elliott and Timmermann (2003))

$$\begin{aligned}
\omega^*_{oth} &= \sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}\mu_{ys_{t+h}} - (\sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}\boldsymbol{\mu}'_{\widehat{\mathbf{y}}s_{t+h}})\boldsymbol{\omega}_{th} \equiv \bar{\mu}_{yt} + \bar{\boldsymbol{\mu}}'_{\widehat{\mathbf{y}}t}\boldsymbol{\omega}_{th}, \\
\omega^*_{th} &= \left(\sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}\left(\boldsymbol{\mu}_{\widehat{\mathbf{y}}s_{t+h}}\boldsymbol{\mu}'_{\widehat{\mathbf{y}}s_{t+h}} + \boldsymbol{\Sigma}_{\widehat{\mathbf{y}}s_{t+h}}\right) - \bar{\boldsymbol{\mu}}_{\widehat{\mathbf{y}}t}\bar{\boldsymbol{\mu}}'_{\widehat{\mathbf{y}}t}\right)^{-1} \\
&\quad \times \left(\sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}(\mu_{ys_{t+h}}\boldsymbol{\mu}_{\widehat{\mathbf{y}}s_{t+h}} + \boldsymbol{\sigma}_{y\widehat{\mathbf{y}}s_{t+h}}) - \bar{\mu}_{yt}\bar{\boldsymbol{\mu}}_{\widehat{\mathbf{y}}t}\right), \qquad (62)
\end{aligned}$$

where $\bar{\mu}_{yt} = \sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}\mu_{ys_{t+h}}$ and $\bar{\boldsymbol{\mu}}_{\widehat{\mathbf{y}}t} = \sum_{s_{t+h}=1}^{n_s} \pi_{s_{t+h},t}\boldsymbol{\mu}_{\widehat{\mathbf{y}}s_{t+h}}$. The standard weights in (8) can readily be obtained by setting $n_s = 1$.

It follows from (62) that the (conditionally) optimal combination weights will vary as the state probabilities vary over time as a function of the arrival of new information, provided that $\mathbf{P}$ is of rank greater than one.

4.2 Nonlinear Combination Schemes

Two types of non-linearities can be admitted in the combination scheme. First, and simplest, non-linear functions of the forecasts can be used in the combination which is nevertheless linear in the unknown parameters:

$$\hat{y}^c_{t+h,t} = \beta_0 + \boldsymbol{\beta}_1 g(\widehat{\mathbf{y}}_{t+h,t}). \qquad (63)$$

Here $g(\widehat{\mathbf{y}}_{t+h,t})$ is a vector-valued function of the set of forecasts that typically includes a lead term that is linear in $\widehat{\mathbf{y}}_{t+h,t}$ in addition to higher order terms similar to a Volterra or Taylor series expansion. The nonlinearity only enters through the shape of the transformation $g(.)$ so the unknown parameters can readily be estimated by OLS although the small-sample properties of such estimates could be an issue. A second and more general method allows for general non-linearity in the unknown combination parameters as well as in the transformation of the raw forecasts, i.e.

$$\hat{y}^c_{t+h,t} = g(\widehat{\mathbf{y}}_{t+h,t}, \boldsymbol{\theta}). \qquad (64)$$

There does not appear to be much work in this area, possibly due to the fact that estimation errors already appear to be large in linear combination schemes and hence can be expected

to be even larger for non-linear combinations whose parameters are generally less robust and more sensitive to outliers than those of the linear schemes.

One exception to this is the paper by Donaldson and Kamstra (1996) which uses artificial neural networks to combine volatility forecasts from underlying models. Their combination scheme takes the form

$$\begin{aligned} \hat{y}^c_{t+h,t} &= \beta_0 + \sum_{j=1}^{n} \beta_j \hat{y}_{t+h,t,j} + \sum_{i=1}^{p} \delta_i g(z_{t+h,t}\gamma_i), \\ g(z_{t+h,t}\gamma_i) &= (1 + \exp(-(\gamma_{0,i} + \gamma_{1,i} z_{1,t} + \gamma_{2,i} z_{2,t})))^{-1} \\ z_{j,t} &= (\hat{y}_{t+h,t,j} - \bar{y}_t)/\hat{\sigma}_{yt}, \\ k &\in \{0, 2\}, \quad p \in \{0, 1, 2, 3\}. \end{aligned} \tag{65}$$

Here $\bar{y}_t$ is the sample estimate of the mean of y, while $\hat{\sigma}_{yt}$ is the standard deviation, using data up to time t. This network uses logistic nodes. The linear model is nested as a special case when $p = 0$ so no nonlinear terms are included. In an out-of-sample forecasting experiment for volatility in daily stork returns, Donaldson and Kamstra find evidence that the neural net combination applied to two underlying forecasts (moving average varaince model and a GARCH(1,1) model) outperforms traditional combination methods.

5 Shrinkage Methods

Shrinkage methods aim to trade off bias in the combination weights against reduced parameter estimation error in estimates of the combination weights. Intuition for how the shrinkage method works is well summarized by Ledoit and Wolf (2004 page 2): "The crux of the method is that those estimated coefficients in the sample covariance matrix that are extremely high tend to contain a lot of positive error and therefore need to be pulled downwards to compensate for that. Similarly, we compensate for the negative error that tends to be embedded inside extremely low estimated coefficients by pulling them upwards." As noted by Ledoit and Wolf (2003), the sample covariance matrix is subject to considerable estimation uncertainty in cases with N large relative to T. This problem can partially be resolved by imposing more structure on the estimator in a way that reduces estimation error

although the key question remains how much and which structure to impose. Shrinkage methods let the forecast combination weights depend on the sample size relative to the number of cross-sectional models to be combined.

Diebold and Pauly (1990) propose to shrink the weights towards equal-weights. Consider the standard linear regression model underlying most forecast combinations

$$\mathbf{Y} = \hat{\mathbf{Y}}\boldsymbol{\omega} + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}), \tag{66}$$

where $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ are $T \times 1$ vectors, $\hat{\mathbf{Y}}$ is the $T \times N$ matrix of forecasts and $\boldsymbol{\omega}$ is the $N \times 1$ vector of combination weights. The standard normal-gamma conjugate prior $\sigma^2 \sim IG(s_0^2, v_0)$, $\boldsymbol{\omega}|\sigma \sim N(\boldsymbol{\omega}_0, \mathbf{M})$ implies that

$$P_0(\boldsymbol{\omega}, \sigma) \propto \sigma^{-N-v_0-1} \exp\left(\frac{-(v_0 s_0^2 + (\boldsymbol{\omega} - \boldsymbol{\omega}_0)'\mathbf{M}(\boldsymbol{\omega} - \boldsymbol{\omega}_0))}{2\sigma^2}\right) \tag{67}$$

Under normality of $\boldsymbol{\varepsilon}$ the likelihood function for the data is

$$L(\boldsymbol{\omega}, \sigma|\mathbf{y}, \hat{\mathbf{y}}) \propto \sigma^{-T} \exp\left(\frac{-(\mathbf{y} - \hat{\mathbf{y}}\boldsymbol{\omega})'(\mathbf{y} - \hat{\mathbf{y}}\boldsymbol{\omega})}{2\sigma^2}\right). \tag{68}$$

These results can be combined to give the marginal posterior for $\boldsymbol{\omega}$ with mean

$$\bar{\boldsymbol{\omega}} = (\mathbf{M} + \hat{\mathbf{y}}'\hat{\mathbf{y}})^{-1}(\mathbf{M}\boldsymbol{\omega}_0 + \hat{\mathbf{y}}'\hat{\mathbf{y}}\hat{\boldsymbol{\omega}}), \tag{69}$$

where $\hat{\boldsymbol{\omega}} = (\hat{\mathbf{y}}'\hat{\mathbf{y}})^{-1}\hat{\mathbf{y}}'\hat{\mathbf{y}}$ is the least squares estimate of $\boldsymbol{\omega}$. Using a prior proportional to $\hat{\mathbf{y}}'\hat{\mathbf{y}}$, namely $\mathbf{M} = g\hat{\mathbf{y}}'\hat{\mathbf{y}}$, we get

$$\bar{\boldsymbol{\omega}} = (g\hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{y}}'\hat{\mathbf{y}})^{-1}(g\hat{\mathbf{y}}'\hat{\mathbf{y}}\boldsymbol{\omega}_0 + \hat{\mathbf{y}}'\hat{\mathbf{y}}\hat{\boldsymbol{\omega}}),$$

which can be used to obtain

$$\bar{\boldsymbol{\omega}} = \boldsymbol{\omega}_0 + \frac{\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0}{1 + g}. \tag{70}$$

Clearly, the larger the value of g, the stronger the shrinkage towards the mean of the prior, $\boldsymbol{\omega}_0$, whereas small values of g suggest putting more weight on the data.

As pointed out by Diebold and Pauly, empirical Bayes methods can alternatively be used to estimate g. Suppose the prior for $\boldsymbol{\omega}$ conditional on σ is Gaussian $N(\boldsymbol{\omega}_0, \tau^2\mathbf{A}^{-1})$. Then

the posterior for $\boldsymbol{\omega}$ is also Gaussian, $N(\bar{\boldsymbol{\omega}}, \boldsymbol{\tau}^{-2}\mathbf{A} + \sigma^{-2}\hat{\mathbf{y}}'\hat{\mathbf{y}})$ and we can replace σ^2 and τ^2 by the estimates

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(\mathbf{y} - \hat{\mathbf{y}}\hat{\boldsymbol{\omega}})'(\mathbf{y} - \hat{\mathbf{y}}\hat{\boldsymbol{\omega}})}{T} \\ \hat{\tau}^2 &= \frac{(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0)'(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0)}{tr(\hat{\mathbf{y}}'\hat{\mathbf{y}})^{-1}} - \hat{\sigma}^2.\end{aligned}$$

This gives rise to an empirical Bayes estimator of $\boldsymbol{\omega}$ whose posterior mean is

$$\bar{\boldsymbol{\omega}} = \boldsymbol{\omega}_0 + \left(\frac{\hat{\tau}^2}{\hat{\sigma}^2 + \hat{\tau}^2}\right)(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0). \tag{71}$$

The empirical Bayes combination shrinks $\hat{\boldsymbol{\omega}}$ towards $\boldsymbol{\omega}_0$ and amounts to setting $g = \hat{\sigma}^2/\hat{\tau}^2$ in (70). Notice that if $\hat{\sigma}^2/\hat{\tau}^2 \to 0$, the OLS estimator is obtained while if $\hat{\sigma}^2/\hat{\tau}^2 \to \infty$, the prior estimate $\boldsymbol{\omega}_0$ is obtained as a special case. Diebold and Pauly argue that the combination weights should be shrunk towards the equal-weighted (simple) average so the combination procedure gives a convex combination of the least-squares and equal weights.

Stock and Watson (2003) also propose shrinkage towards the arithmetic average of forecasts. Let $\hat{\omega}_{it}$ be the least-squares estimator of the weight on the ith model in the forecast combination. The combination weights considered by Stock and Watson take the form

$$\begin{aligned}\omega_{iT} &= \psi\hat{\omega}_{iT} + (1 - \psi)(1/N), \\ \psi &= \max(0, 1 - \kappa N/(T - h - N - 1)),\end{aligned}$$

where κ regulates the strength of the shrinkage. Stock and Watson consider values $\kappa = 1/4, 1/2$ or 1. As the sample size, T, rises relative to N, the least squares estimate gets a larger weight. Indeed, if T grows at a faster rate than N, the least squares weight will, in the limit, get a weight of unity.

5.1 Applications to Portfolio Analysis

Further insights can be gained by noting that the problem of forming mean-variance efficient portfolios in finance is mathematically equivalent to that of combining forecasts, c.f. the papers in Dunis, Timmermann and Moody (2001). In finance, the standard optimization problem minimizes the portfolio variance $\boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega}$ subject to a given portfolio return, $\boldsymbol{\omega}'\boldsymbol{\mu} = \mu_0$,

where $\boldsymbol{\mu}$ is a vector of mean returns while $\boldsymbol{\Sigma}$ is the covariance matrix of asset returns. Imposing also the constraint that the portfolio weights sum to unity, we have

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \{\boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega}\} \\ &s.t. \;\; \boldsymbol{\omega}'\boldsymbol{\iota} = 1, \\ &\boldsymbol{\omega}'\boldsymbol{\mu} = \mu_0 \; . \end{aligned} \tag{72}$$

This problem has the solution

$$\boldsymbol{\omega}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} \; \boldsymbol{\iota}) \left[(\boldsymbol{\mu} \; \boldsymbol{\iota})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} \; \boldsymbol{\iota}) \right]^{-1} \begin{pmatrix} \mu_0 \\ 1 \end{pmatrix}, \tag{73}$$

The only difference to the optimal solution from the forecast combination problem is that a minimum variance portfolio is derived for each separate value of the mean portfolio return, μ_0, whereas in the forecast combination problem the constraint $\boldsymbol{\omega}'\boldsymbol{\iota} = 1$ is generally interpreted as guaranteeing an unbiased combined forecast−assuming of course that the individual forecasts are also unbiased.

In a portfolio application, Ledoit and Wolfe (2003) propose instead to shrink towards a point implied by a single factor structure common in finance. Suppose that the individual forecast errors are affected by a single common factor, F_{et}

$$e_{it} = \alpha_i + \beta_i F_{et} + \varepsilon_{it}. \tag{74}$$

where the idiosyncratic residuals, ε_{it}, are assumed to be orthogonal across forecasting models and uncorrelated with F_{et}. This single factor model has a long tradition in finance but is also a natural starting point for forecasting purposes since forecast errors are generally strongly positively correlated. One could use principal components techniques to extract the common factor from the forecast errors. Letting $\sigma^2_{f_e}$ be the variance of F_{et}, the covariance matrix of the forecast errors becomes

$$\boldsymbol{\Sigma}_{ef} = \sigma^2_{f_e} \boldsymbol{\beta}\boldsymbol{\beta}' + \mathbf{D}_{\varepsilon}, \tag{75}$$

where $\boldsymbol{\beta} = (\beta_1 \cdots \beta_N)'$ is the vector of factor sensitivities, while $\mathbf{D}_{\varepsilon}$ is a diagonal matrix with the individual values of $Var(\varepsilon_{it})$ on the diagonal. Estimation of $\boldsymbol{\Sigma}_{ef}$ requires determining

only $2N+1$ parameters. Consistent estimates of these parameters are easily obtained by estimating (74) by OLS, equation by equation, to get

$$\hat{\mathbf{\Sigma}}_{ef} = \hat{\sigma}^2_{f_e}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \hat{\mathbf{D}}_{\varepsilon}.$$

Typically this covariance matrix is biased due to the stringent assumption that $\mathbf{D}_{\varepsilon}$ is diagonal. For example, there may be more than a single common factor in the forecast errors and some forecasts may omit the same relevant variable in which case blocks of forecast errors will be correlated. Though biased, the single factor covariance matrix is typically surrounded by considerably smaller estimation errors than the unconstrained matrix, $E[\mathbf{ee}']$, which is usually estimated by

$$\hat{\mathbf{\Sigma}}_e = \frac{1}{T-h}\sum_{\tau=h}^{T}\mathbf{e}_{\tau,\tau-h}\mathbf{e}'_{\tau,\tau-h},$$

where $\mathbf{e}_{\tau,\tau-h}$ is an $N \times 1$ matrix of forecast errors. This estimator requires estimating $N(N+1)/2$ parameters. Using $\hat{\mathbf{\Sigma}}_{ef}$ as the shrinkage point, Ledoit and Wolf (2003) propose minimizing the following quadratic loss as a function of the shrinkage parameter, α,

$$L(\alpha) = ||\alpha\hat{\mathbf{\Sigma}}_{ef} + (1-\alpha)\hat{\mathbf{\Sigma}}_e - \mathbf{\Sigma}_e||^2,$$

where $||.||^2$ is the Frobenius norm, i.e. $||\mathbf{Z}||^2 = trace(\mathbf{Z}^2)$, $\hat{\mathbf{\Sigma}}_e$=$(1/T)\mathbf{e}(\mathbf{I} - \boldsymbol{\iota}\boldsymbol{\iota}'/T)\mathbf{e}'$ is the sample covariance matrix and $\mathbf{\Sigma}_e$ is the true matrix of squared forecast errors, $E[\mathbf{e}'\mathbf{e}]$, where $\mathbf{e}$ is a $T \times N$ matrix of forecast errors . They demonstrate that the optimal shrinkage satisfies

$$\alpha^* = \frac{1}{T}\frac{\pi - \rho}{\gamma} + O(\frac{1}{T^2}),$$

where

$$\begin{aligned}
\pi &= \sum_{i=1}^{N}\sum_{j=1}^{N} AsyVar(\sqrt{T}\hat{\sigma}_{ij}), \\
\rho &= \sum_{i=1}^{N}\sum_{j=1}^{N} AsyCov(\sqrt{T}\hat{f}_{ij}, \sqrt{T}\hat{\sigma}_{ij}), \\
\gamma &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\phi_{ij} - \sigma_{ij})^2.
\end{aligned}$$

$\hat{f}_{ij}$ is the (i,j) entry of $\hat{\mathbf{\Sigma}}_{ef}$, $\hat{\sigma}_{ij}$ is the (i,j) element of $\hat{\mathbf{\Sigma}}_e$ and ϕ_{ij} is the (i,j) element of the single factor covariance matrix, $\mathbf{\Sigma}_{ef}$, while σ_{ij} is the (i,j) element of $\mathbf{\Sigma}_e$. Hence, π measures

the (scaled) sum of asymptotic variances of the sample covariance matrix ($\hat{\mathbf{\Sigma}}_e$), p measures the (scaled) sum of asymptotic covariances of the single-factor covariance matrix ($\hat{\mathbf{\Sigma}}_{ef}$), while γ measures the degree of misspecification (bias) in the single factor model. Ledoit and Wolf propose consistent estimators for the moments $\hat{\pi}, \hat{\rho}$ and $\hat{\gamma}$ under the assumption of IID forecast errors.[7]

Provided that a set of weak regularity conditions hold for the forecast errors, Ledoit and Wolf show that the following estimator is consistent for the optimal shrinkage constant

$$\hat{\kappa} = \frac{\hat{\pi} - \hat{\rho}}{\hat{\gamma}},$$

The proposed shrinkage estimator, $\hat{\mathbf{S}}$, takes the form

$$\hat{\mathbf{S}} = \frac{\hat{\kappa}}{T}\hat{\mathbf{\Sigma}}_{ef} + (1 - \frac{\hat{\kappa}}{T})\hat{\mathbf{\Sigma}}_e. \tag{76}$$

In practice one can bound the shrinkage estimator to lie in the zero-one interval:

$$\frac{\hat{\kappa}}{T} = \max(0, \min(\frac{\hat{\kappa}}{T}, 1)). \tag{77}$$

A closely related combination scheme proposed by Bunn (1985) lets the forecasting method depend on the sample size. Suppose we have a fixed number of forecasts, N, to combine while the sample size, T, is expanding. The relative performance of a given forecast combination scheme depends on the sample size at hand, with more sophisticated covariance-based combination methods requiring a larger sample size to estimate the parameters sufficiently precisely. One possibility is then to initially use equal weights when the sample size is small and parameter estimation errors a big problem, switch to weights that are inversely proportional to the forecasts' relative MSE-values as the sample size grows (which corresponds to only estimating the diagonal elements of the covariance matrix of forecast errors), followed by a combination based on the full covariance matrix when a large sample size (relative to N) becomes available.

[7] It is worth pointing out that the assumption that $\mathbf{e}$ is IID is unlikely to hold for forecast errors which could share common dynamics in first, second or higher order moments or even be serially correlated, c.f. Diebold (1988).

5.2 Portfolio Weight Constraints

Shrinkage has a very interesting relationship with portfolio weight constraints in finance. It is commonplace to consider minimization of portfolio variance subject to a set of equality and inequality constraints on the portfolio weights. Portfolio weights are often constrained to be non-negative (no short selling) and not to exceed certain upper bounds. Consider the optimization program

$$\begin{aligned} \boldsymbol{\omega}^* &= \arg\min_{\boldsymbol{\omega}} \frac{1}{2}\boldsymbol{\omega}'\hat{\Sigma}\boldsymbol{\omega} \\ s.t. \quad \boldsymbol{\omega}'\boldsymbol{\iota} &= 1 \\ \omega_i &\geq 0,\, i = 1, ..., N \\ \omega_i &\leq \bar{\omega},\, i = 1, ..., N \end{aligned} \tag{78}$$

with the associated Kuhn-Tucker conditions:

$$\begin{aligned} \sum_j \hat{\Sigma}[i,j]\omega_j - \lambda_i + \delta_i &= \lambda_0 \geq 0 \quad i = 1, ..., N \\ \lambda_i &\geq 0 \quad \text{and } \lambda_i = 0 \text{ if } \omega_i > 0 \\ \delta_i &\geq 0 \quad \text{and } \delta_i = 0 \text{ if } \omega_i < \bar{\omega} \end{aligned}$$

where $\hat{\Sigma}$ is an estimate of the covariance matrix for some cross-section of asset returns with row i, column j element $\hat{\Sigma}[i,j]$. Lagrange multipliers for the lower and upper bounds are collected in the vectors $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)'$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_N)'$; λ_0 is the Lagrange multiplier for the constraint that the weights sum to one.

Constraints on portfolio weights effectively have two effects. First, they shrink the largest elements of the covariance matrix towards zero. This reduces the effects of estimation error that can be expected to be strongest for assets with extreme weights. The second effect is that it may introduce specification errors to the extent that the true population values of the optimal weights actually lie outside the assumed interval.

Jagannathan and Ma (2003) show the following interesting result. Let

$$\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} + (\boldsymbol{\delta}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\delta}') - (\boldsymbol{\lambda}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\lambda}'). \tag{79}$$

Then $\tilde{\boldsymbol{\Sigma}}$ is symmetric and positive semi-definite, and the solution to the constrained portfolio variance minimization problem (78) is one of its global minimum variance portfolios. This

shows that constructing a constrained global minimum variance portfolio is equivalent to constructing an unconstrained minimum variance portfolio based on the modified covariance matrix $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} + (\boldsymbol{\delta}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\delta}') - (\boldsymbol{\lambda}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\lambda}')$.

Furthermore, it turns out that $\tilde{\boldsymbol{\Sigma}}$ can be interpreted as a shrinkage version of $\hat{\boldsymbol{\Sigma}}$. To see this, consider the weights that are affected by the lower bound and for which $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} - (\boldsymbol{\lambda}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\lambda}')$. When the constraint for the lower bound is binding (so a weight would have been negative), the covariances of a particular forecast error with all other errors are reduced by the strictly positive Lagrange multipliers and its variance is shrunk. Imposing the non-negativity constraints shrinks the largest covariances that would have resulted in negative weights. Furthermore, since the largest estimates of the covariance are more likely to be the result of estimation error, such shrinkage can have the effect of reducing the estimation error.

In the case of the upper bounds, those forecasts whose unconstrained weights would have exceeded $\bar{\omega}$ are also the ones for which the variance and covariance estimates tend to be smallest. These forecasts have strictly positive Lagrange multipliers on the upper bound constraint, meaning that their forecast error variance in the modified covariance matrix $\tilde{\boldsymbol{\Sigma}}$ will be increased by $2\delta_i$ whereas their covariances with other forecast errors increases by $\delta_i + \delta_j$. Again this corresponds to shrinkage towards the cross-sectional average of the variances and covariances.

6 Empirical Evidence

The empirical literature on forecast combinations is voluminous and includes work in several areas in management science, economics, operations research, meteorology, psychology and finance. The work in economics dates back to Reid (1968) and Bates and Granger (1969). Although details and results vary across studies, it is nevertheless possible to draw some broad conclusions from much of this work. Such conclusions must necessarily come with a stronger than usual caveat emptor since for each point it is possible to construct counter examples. This is necessarily the case since findings depend on the number of models, N, (as well as their type), the sample size, T, the extent of instability in the underlying data

set and the structure of the covariance matrix of the forecast errors (e.g., diagonal or with similar correlations).

Nevertheless, empirical findings in the literature on forecast combinations broadly suggest that (i) parameter estimation error is often very important in explaining the performance of various combination schemes. Methods aimed at reducing such errors (such as shrinkage or combination methods that ignore correlations between forecasts) tend to perform well; (ii) trimming of the worst models and clustering of models with similar forecasting performance prior to combination can yield large improvements, especially in situations involving large numbers of forecasts; and (iii) some time-variation or adaptive adjustment in the combination weights (or perhaps in the underlying models being combined) can also improve forecasting performance.

6.1 Simple Combination Schemes are hard to beat

It has often been found that simple combinations−that is, combinations that do not require estimating many parameters such as arithmetic averages or weights based on the inverse mean squared forecast error−do better than more sophisticated rules relying on estimating optimal weights that depend on the full variance-covariance matrix of forecast errors, c.f. Bunn (1985), Clemen and Winkler (1986), Dunis, Laws and Chauvin (2001), Figlewski and Urich (1983) and Makridakis et al (1982, 1983).

Palm and Zellner (1992, p. 699) concisely summarize the advantages of a simple average forecast:

"1. Its weights are known and do not have to be estimated, an important advantage if there is little evidence on the performance of individual forecasts or if the parameters of the model generating the forecasts are time-varying;

2. In many situations a simple average of forecasts will achieve a substantial reduction in variance and bias through averaging out individual bias;

3. It will often dominate, in terms of MSE, forecasts based on optimal weighting if proper account is taken of the effect of sampling errors and model uncertainty on the estimates of the weights."

Despite the impressive empirical track record of equal-weighted forecast combinations we stress that the theoretical justification for this method critically depends on the ratio of forecast error variances not being too far away from unity and also depends on the correlation between forecast errors not varying too much across pairs of models. Consistent with this, Gupta and Wilson (1987) find that the performance of equal weighted combinations depends strongly on the relative size of the variance of the forecast errors associated with different methods. When these are similar, equal weights perform well, while when larger differences are observed, differential weighting of forecasts is generally required.

The importance of instability in the combination weights has been recognized by, inter alia, Clemen and Winkler (1986), Kang (1986) and - in the context of non-stationarities - Diebold and Pauly (1987). If the instability is sufficiently important to render precise estimation of combination weights nearly impossible, equal-weighting of forecasts may become an attractive alternative as pointed out by Figlewski and Urich (1983) and Palm and Zellner (1992).

In one of the most comprehensive studies to date, Stock and Watson (2001) consider combinations of a range of linear and nonlinear models fitted to a very large set of US macroeconomic variables. They find strong evidence in support of using forecast combination methods, particularly the average or median forecast and the forecasts weighted by their inverse MSE. The overall dominance of the combination forecasts holds at the one, six and twelve month horizons. Furthermore, the best combination methods combine forecasts across many different time-series models.

Similarly, in a time-series simulation experiment, Winkler and Makridakis (1983) find that a weighted average with weights inversely proportional to the sum of squared errors or a weighted average with weights that depend on the exponentially discounted sum of squared errors perform better than the best individual forecasting model, equal-weighting or methods that require estimation of the full covariance matrix for the forecast errors.

Results regarding the performance of equal-weighted forecast combinations may be sensitive to the loss function underlying the problem. Elliott and Timmermann (2003) find in an empirical application that the optimal weights in a combination of inflation survey

forecasts and forecasts from a simple autoregressive model strongly depend on the degree of asymmetry in the loss function. In the absence of loss asymmetry, the autoregressive forecast does not add much information. However, under asymmetric loss (in either direction), both sets of forecasts appear to contain information and have non-zero weights in the combined forecast. Their application confirms the frequent finding that equal-weights outperform estimated optimal weights under MSE loss. However, it also shows very clearly that this result can be overturned under asymmetric loss where use of estimated optimal weights may lead to much smaller average losses out-of-sample.

6.2 Choosing the forecast with the best track record is often a bad idea

Many studies have found that combination dominates the best individual forecast in out-of-sample forecasting experiments. For example, Makridakis et al (1982) report that a simple average of six forecasting methods performed better than the underlying individual forecasts. In simulation experiments Gupta and Wilson (1987) also find combination superior to the single best forecast. Makridakis and Winkler (1983) report large gains from simply averaging forecasts from individual models over the performance of the best model. Hendry and Clements (2002) explain the better performance of combination methods over the best individual model by misspecification of the models caused by deterministic shifts in the underlying data generating process. Naturally, the models cannot be misspecified in the same way with regard to this source of change, or else diversification gains would be zero.

Aiolfi and Timmermann (2004) find evidence of persistence in the out-of-sample performance of linear and non-linear forecasting models fitted to a large set of macroeconomic time-series in the G7 countries. Models that were in the top and bottom quartiles when ranked by their historical forecasting performance have a higher than average chance of remaining in the top and bottom quartiles, respectively, in the out-of-sample period. They also find systematic evidence of 'crossings', where the previous best models become the worst models in the future or vice versa, particularly among the linear forecasting models. This appears to be more of a problem for linear than for nonlinear forecasting models. They also

find that many forecast combinations produce lower out-of-sample MSE than a strategy of selecting the previous best forecasting model irrespective of the length of the backward-looking window used to measure past forecasting performance.

6.3 Trimming of the worst models is often required

Trimming of forecasts can occur at two levels. First, it can be adopted as a form of outlier reduction rule (c.f. Chan, Stock and Watson (1999)) at the initial stage that produces forecasts from the individual models. Second it can be used in the combination stage where models deemed to be too poor may simply be discarded. Since the first form of trimming has more to do with specification of the individual models underlying the forecast combination, we concentrate on the latter form of trimming which has been used successfully in many studies. Most obviously, when many forecasts get a weight close to zero, improvements due to reduced parameter estimation errors can be gained by dropping such models.

Winkler and Makridakis (1983) find that including very poor models in an equal-weighted combination can substantially worsen forecasting performance. Stock and Watson (2003) also find that the simplest forecast combination methods such as trimmed equal weights and slowly moving weights perform best and that such combinations do better than forecasts from a dynamic factor model.

In their thick modeling approach, Granger and Jeon (2004) recommend trimming, say, the five or ten percent of the worst models, although the extent of the trimming will depend on the application at hand.

More aggressive trimming has also been proposed. In a forecasting experiment involving the prediction of stock returns by means of a large set of forecasting models, Aiolfi and Favero (2003) investigate the performance of a large set of trimming schemes. Their findings indicate that the best performance is obtained when the top 20% of the forecasting models is combined in the forecast so that 80% of the models (ranked by their R^2-value) are trimmed.

6.4 Shrinkage often improves performance

By and large shrinkage methods have performed quite well in empirical studies. In an empirical exercise containing four real-time forecasts of nominal and real GNP, Diebold and Pauly (1990) report that shrinkage weights systematically improve upon the forecasting performance over methods that select a single forecast or use least squares estimates of the combination weights. They direct the shrinkage towards a prior reflecting equal weights and find that the optimal degree of shrinkage tends to be large. Similarly, Stock and Watson (2003) find that shrinkage methods perform best when the degree of shrinkage (towards equal weights) is quite strong.

Aiolfi and Timmermann (2004) explore persistence in the performance of forecasting models by proposing a set of combination strategies that first pre-select models into either quartiles or clusters on the basis of the distribution of past forecasting performance across models, pool forecasts within each cluster and then estimate optimal combination weights that are shrunk towards equal weights. These conditional combination strategies lead to better average forecasting performance than simpler strategies commonly used such as using the historical best model or simply averaging across all forecasting models or a small subset of these.

Elliott (2004) undertakes a simulation experiment where he finds that although shrinkage methods always dominate least squares estimates of the combination weights, the performance of the shrinkage method can be quite sensitive to the shrinkage parameter and that none of the standard methods for determining this parameter work particularly well.

Given the similarity of the mean-variance optimization problem in finance to the forecast combination problem, it is not surprising that empirical findings in finance mirror those in the forecast combination literature. For example, it has generally been found in applications to asset returns that sample estimates of portfolio weights solving the mean-variance problem (72) are extremely sensitive to small changes in sample means. In addition they are highly sensitive to variations in the inverse of the covariance matrix estimate, $\hat{\mathbf{\Sigma}}^{-1}$.

Jobson and Korkie (1980) show that the sample estimate of the optimal portfolio weights can be characterized as the ratio of two (random) estimators, each of whose first and second

moments can be derived in closed form. They use Taylor series expansions to derive an approximate solution for the first two moments of the optimal weights, noting that higher order moments can be characterized under additional normality assumptions. They also derive the asymptotic distribution of the portfolio weights for the case where N is fixed and T goes to infinity. In simulation experiments they demonstrate that the sample estimates of the portfolio weights are highly volatile and can take extreme values that lead to poor out-of-sample performance.

It is widely recognized in finance that imposing portfolio weight constraints generally leads to improved out-of-sample performance of mean-variance efficient portfolios. For example, Jagannathan and Ma (2003) find empirically that once such constraints on portfolio weights are accounted for, other refinements of covariance matrix estimation have little additional effect on the variance of the optimal portfolio. Since they also demonstrate that portfolio weight constraints can be interpreted as a form of shrinkage, these findings lend support to using shrinkage methods as well.

In the context of portfolio formation Ledoit and Wolf (2003) report that the out-of-sample standard deviation of portfolio returns based on a shrunk covariance matrix is significantly lower than the standard deviation of portfolio returns based on more conventional estimates of the covariance matrix.

Notice that shrinkage and trimming tend to work in opposite directions - at least if the shrinkage is towards equal weights. Shrinkage tends to give similar weights to all models whereas trimming completely discards a subset of models. If some models produce extremely poor out-of-sample forecasts, shrinkage can be expected to perform poorly if the combined forecast is shrunk too aggressively towards an equal-weighted average. For this reason, shrinkage preceded by a trimming step may work well in many situations.

6.5 Limited time-variation in the combination weights may be helpful

The evidence on the value of allowing for time-varying combinations in the combination weights is somewhat mixed. Time-variations in forecasts can be introduced either in the

individual models underlying the combination or in the combination weights themselves and both approaches have been considered. The idea of time-varying forecast combinations goes back to the advent of the combination literature in economics. Bates and Granger (1969) used combination weights that were adaptively updated as did many subsequent studies such as Winkler and Makridakis (1983). Newbold and Granger (1974) considered values of the window length, v, in (48) and (49) between 1 and 12 and values of the discounting factor, λ, in (51) and (52) between 1 and 2.5. Their results suggested that there is an interior optimum around $v = 6, \alpha = 0.5$ where the adaptive updating method (50) does best whereas the rolling window combinations generally do best for the longest windows, i.e., $v = 9$ or $v = 12$, and the best exponential discounting was around 2 or 2.5. This is consistent with the finding by Bates and Granger (1969) that high values of the discounting factor tend to work best. A method that combines a Holt-Winters and stepwise autoregressive forecast was found to perform particularly well. Winkler and Makridakis (1983) report similar results and also find that the longer windows, v, in equations such as (48) and (49) tend to produce the most accurate forecasts, although the best results among the discounting methods were found for relatively low values of the discount factor.

In a combination of forecasts from the Survey of Professional Forecasters and forecasts from simple autoregressive models applied to six macroeconomic variables, Elliott and Timmermann (2003) investigate the out-of-sample forecasting performance produced by different constant and time-varying forecasting schemes such as (60). Compared to a range of other time-varying forecast combination methods, a two-state regime switching method produces a lower MSFE in four (against rolling window and Kalman filter forecasts) or five (against a forecast of the type proposed by Deutsch, Granger and Terasvirta (1994), (55)) out of six cases. They argue that the evidence suggests that the best forecast combination method allows the combination weights to vary over time but in a mean-reverting manner. Unsurprisingly, allowing for three states leads to worse forecasting performance for four of the six variables under consideration.

Stock and Watson (2003) report that the combined forecasts that do best in their study are the time-varying parameter (TVP) forecast with very little time variation, the simple

mean and a trimmed mean. They conclude that "the results for the methods designed to handle time variation are mixed. The TVP forecasts sometimes work well but sometimes work quite poorly and in this sense are not robust; the larger the amount of time variation, the less robust are the forecasts. Similarly, the discounted MSE forecasts with the most discounting.... are typically no better than, and sometimes worse than, their counterparts with less or no discounting."

This leads Stock and Watson (2003) to conclude "This "forecast combination puzzle" - the repeated finding that simple combination forecasts outperform sophisticated adaptive combination methods in empirical applications - is, we think, more likely to be understood in the context of a model in which there is widespread instability in the performance of individual forecast, but the instability is sufficiently idiosyncratic that the combination of these individually unstably performing forecasts can itself be stable."

6.6 Monte Carlo Simulations

To illustrate the factors that determine the precision of the various forecast methodologies, we conduct a Monte Carlo experiment in the context of the earlier factor model assumed in (41). This setting is appropriate for our purpose for the following reasons. First, factor models are widely used empirically to forecast macroeconomic and financial time series. Second, intuition can be gained in terms of distribution of factor loadings/exposures and variability of the individual factors.

We assume the following dynamic model for the n_f orthogonalized factors

$$\mathbf{F}_t = \mathbf{B}_F \mathbf{F}_{t-1} + \boldsymbol{\varepsilon}_{F_t}, \tag{80}$$

where $\boldsymbol{\varepsilon}_{F_t} \sim N(0, \mathbf{D}_{\varepsilon_F})$ while $\mathbf{D}_{\varepsilon_F}$ is an $n_f \times n_f$ diagonal matrix with entries

$$\mathbf{D}_{\varepsilon_F} = \begin{pmatrix} \sigma^2_{F_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{F_2} & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & & \sigma^2_{F_{n_f}} \end{pmatrix}.$$

We assume that the eigenvalues of $\mathbf{B}_F$ all lie outside the unit circle so $(\mathbf{I}-\mathbf{B}_F)^{-1}$ exists and the initial value $\mathbf{F}_0$ can be drawn from the unconditional distribution of the factors. The target variable and the individual forecasts are assumed to have factor structure

$$\begin{aligned} Y_{t+1} &= \mu_y + \boldsymbol{\beta}_{yF}'\mathbf{F}_{t+1} + \varepsilon_{yt+1}, \quad \varepsilon_{yt+1} \sim N(0, \sigma^2_{\varepsilon_Y}) \\ \hat{Y}_{it+1} &= \mu_i + \boldsymbol{\beta}_{iF}'\mathbf{F}_{t+1} + \varepsilon_{it+1}, \quad \varepsilon_{it+1} \sim N(0, \sigma^2_{\varepsilon_i}), \quad i = 1, ..., N. \end{aligned}$$

where $E(\varepsilon_{it+1}\varepsilon_{jt+1}) = 0$ if $i \neq j$, $E[\varepsilon_{it+1}\varepsilon_{yt+1}] = 0$, $E[\varepsilon_{yt+1}\boldsymbol{\varepsilon}_{Ft}] = E[\varepsilon_{it+1}\boldsymbol{\varepsilon}_{Ft}] = \mathbf{0}$, for $i = 1, ..., N$.

This gives the following convenient form of the (unconditional) covariance-matrix in (7):

$$\begin{aligned} \sigma^2_y &= \boldsymbol{\beta}_{yF}'(\mathbf{I}-\mathbf{B}_F^2)^{-1}\mathbf{D}_{\varepsilon_F}\boldsymbol{\beta}_{yF} + \sigma^2_{\varepsilon_Y}, \\ \boldsymbol{\sigma}_{y\hat{\mathbf{y}}}[i] &= \boldsymbol{\beta}_{iF}'(\mathbf{I}-\mathbf{B}_F^2)^{-1}\mathbf{D}_{\varepsilon_F}\boldsymbol{\beta}_{yF} \\ \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}[i,j] &= \boldsymbol{\beta}_{iF}'(\mathbf{I}-\mathbf{B}_F^2)^{-1}\mathbf{D}_{\varepsilon_F}\boldsymbol{\beta}_{jF} + I_{\{i=j\}}\sigma^2_{\varepsilon_i}, \end{aligned} \tag{81}$$

where $I_{\{i=j\}}$ is an indicator function that equals unity if $i = j$, otherwise is zero.

This factor model is quite general so we impose additional structure to ensure that the individual forecasts are sensible. In particular, notice that the population value of the best linear predictor from a regression of Y_t on $\hat{Y}_{it}$ (without an intercept) is simply

$$\frac{\boldsymbol{\beta}_{iF}'(\mathbf{I}-\mathbf{B}_F^2)^{-1}\mathbf{D}_{\varepsilon_F}\boldsymbol{\beta}_{yF}}{\boldsymbol{\beta}_{iF}'(\mathbf{I}-\mathbf{B}_F^2)^{-1}\mathbf{D}_{\varepsilon_F}\boldsymbol{\beta}_{iF}' + \sigma^2_{\varepsilon_i}}. \tag{82}$$

We can choose parameter values such that this is equal to unity, ensuring that the individual forecasts are unbiased. This is the a sensible assumption that is often made in the literature.

Furthermore, if $\boldsymbol{\beta}_{iF} = \boldsymbol{\beta}_{jF}$ and $\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon j}$ for all i, j, then all diagonal elements of $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ are identical and the off-diagonal elements are also the same. This corresponds to the earlier case where equal weights (not necessarily summing to one) are optimal. Letting $\sigma^2_{F_i} = \sigma^2_F$ for $i = 1, .., n_f$, $\sigma^2_{\varepsilon_i} = \sigma^2_\varepsilon$ for $i = 1, .., N$, assuming $\mathbf{B}_F = \mathbf{O}$ and letting $\boldsymbol{\beta}_{iF} = \beta\boldsymbol{\iota}$, $\boldsymbol{\beta}_{yF} = \beta_y\boldsymbol{\iota}$, we have

$$\boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{\sigma^2_\varepsilon}\left(\mathbf{I} - \frac{Nn_f\beta^2\sigma^2_F}{\sigma^2_\varepsilon + Nn_f\beta^2\sigma^2_F}\boldsymbol{\iota}\boldsymbol{\iota}'\right),$$

so that

$$\begin{aligned}\omega^* &= \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}} \\ &= \frac{1}{\sigma_\varepsilon^2}\begin{pmatrix} 1-\frac{Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} & \frac{-Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} & \cdots \\ \frac{-Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} & 1-\frac{Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} & \cdots \\ \vdots & & \ddots \\ \frac{-Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} & & 1-\frac{Nn_f\beta^2\sigma_F^2}{\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2} \end{pmatrix}\begin{pmatrix} n_f\beta\beta_y\sigma_F^2 \\ \vdots \\ \\ n_f\beta\beta_y\sigma_F^2 \end{pmatrix},\end{aligned}$$

It can be seen that equal weights that sum to unity are optimal provided that

$$\frac{n_f\beta\beta_y\sigma_\varepsilon^2\sigma_F^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2+Nn_f\beta^2\sigma_F^2)}=\frac{1}{N}.$$

This only holds if

$$\sigma_\varepsilon^2 = n_f N\beta\sigma_F^2(\beta_y-\beta). \tag{83}$$

It is clear from this, however, that the case where the combination weights are identical *and* sum to unity is a very special case and that restricting the regression coefficients on the individual forecasts to be unity in the univariate regressions via (82) does not ensure that it is optimal to use equal weights. Variations in the variance-covariance parameters across forecasts introduce heterogeneity in forecasting performance and are likely to make equal-weighted forecasts sub-optimal even in population.

We are now ready to set up the Monte Carlo experiments. In all experiments we use two factors, i.e. $n_f = 2$, so that $F = 1, 2$. The first six experiments assume that forecasts are unbiased and set $\mu_y = \mu_i = 0$ $(i = 1, ..., N)$. More specifically, we vary T from 100 to 500 and 1000 and vary N from 4 to 10 and 20. All forecasts are 'pseudo out of sample' and hence are computed based on recursive parameter estimates using only information available at the time of the forecast.

In the base experiment (experiment 1) we assume that β_{i1} solves (83) for all i so that the optimal weights are identical and sum to unity in population. Furthermore, we set

$$\begin{aligned}\boldsymbol{\beta}_y &= (1 \;\; 1)' \\ \sigma_{\varepsilon_Y} &= \sigma_{\varepsilon_{F_1}} = \sigma_{\varepsilon_{F_2}} = 1 \\ \sigma_{\varepsilon_i} &= 1 \quad i = 1, .., N \\ \mathbf{B}_F &= \mathbf{0}.\end{aligned}$$

In experiments 2-7 we assume that $\beta_{i1} = 0.5$, $(i = 1, ..., N)$ while β_{i2} is set to solve (82) for $i = 1, ..., N$ ensuring that the regression coefficient of Y_{t+1} on an individual forecast $\hat{Y}_{it+1}$ is unity. Heterogeneity in the factor loadings is introduced by drawing the factor loadings, β_{if}, from a Beta distribution centered on 0.5 with either low dispersion (coresponding to a $Beta(5,5)$ distribution) or high dispersion (corresponding to a $Beta(1,1)$ distribution).

We alter the base scenario as follows:

Scenarios	change in parameters
1 base scenario ($\boldsymbol{\omega}'\boldsymbol{\iota} = 1$)	$-$
2 identical weights ($\boldsymbol{\omega}'\boldsymbol{\iota} \neq 1$)	β_{i2} solves (82)
3 dynamics	$\mathbf{B}_F = 0.9 \times \mathbf{I}$
4 weak heterogeneity	$\beta_{if} \sim Beta(5,5)$
5 strong heterogeneity	$\beta_{if} \sim Beta(1,1)$
6 factor-loadings in blocks	$\boldsymbol{\beta}'_{i1} = \begin{cases} 1 & \text{if } 1 \leq i \leq N/2 \\ 0 & \text{if } N/2 < i \leq N \end{cases}$ $\boldsymbol{\beta}'_{i2} = \begin{cases} 0 & \text{if } 1 \leq i \leq N/2 \\ 1 & \text{if } N/2 < i \leq N \end{cases}$
7 biased forecast	$\mu_i = \begin{cases} 1/2 & \text{if } 1 \leq i \leq N/2 \\ 0 & \text{if } N/2 < i \leq N \end{cases}$

We compare the following nine combination methods:

(i) GR1 : Unconstrained OLS (33)(i)

(ii) GR2 : OLS w/o constant (33)(ii)

(iii) GR3 : Constrained OLS w/o constant (33)(iii)

(iv) BEST : Forecast from previous best model

(v) EW : Simple equal weighted forecast

(vi) PEW : Projection on constant, EW forecast

(vii) Shrink1 : Shrinkage with $\kappa = 0.25$

(viii) Shrink2 : Shrinkage with $\kappa = 0.5$

(ix) Shrink3 : Shrinkage with $\kappa = 1$

Results are reported in Table 1 in the form of out-of-sample MSE values relative to the MSE generated by the unconstrained OLS method proposed by Granger and Ramanathan (1984). The following conclusions are obtained. In the base scenario the best combination scheme among those proposed by Granger and Ramanathan (1984) is to exclude an intercept and impose that the weights sum to unity. This holds in population and across all sample sizes and cross-sections: imposing a true constraint ensures efficiency gains. The improvement over the most general OLS regression (33)(i) is, however, quite marginal - about 1-2%. When the true weights do not sum to unity, as in the second scenario, the combination that imposes the summability constraint produces MSE performance that is frequently much worse than the unconstrained model. Constraining the intercept to be zero always leads to better performance than under the unconstrained benchmark model when this constraint holds as in experiments 2-6.

Choosing the single best model does not lead to good forecasting performance in the experiments without heterogeneity where (by construction) the forecasting models are equally good. Combining models is therefore a good idea in such experiments as it allows the forecaster to dilute the noise in the individual forecasts, ε_{it}. Turning to experiments 4 and 5, it is clear that the relative out-of-sample performance of the previous best model improves as the degree of heterogeneity across models gets stronger. The one case where it beats the benchmark model is when $N = 20$, $T = 100$ so the estimated combination weights are associated with large errors.

In the base scenario by construction the simple equal-weighted average of forecasts performs best since it imposes a true constraint on the combination weights. However, the simple equal-weighted forecast is not producing good forecasts in the other scenarios even though the parameters of the Monte Carlo are chosen such that the population value of a regression of Y_{t+1} on the individual forecasts $\hat{Y}_{it+1}$ is unity. The reason is that although equal weights are optimal in this setting, they need not sum up to unity. For this reason we consider a simple scheme that regresses Y_{t+1} on an intercept and the equal-weighted forecast, $\bar{Y}_{t+1} = (1/N)\sum_{i=1}^{N} \hat{Y}_{it+1}$. It is clear that this combination leads to much better results and in fact does best among all combination schemes in experiments 2-7 when N is large so

reduction in parameter estimation error becomes important.

Turning finally to the shrinkage forecasts, it is clear that these generally improve on the benchmark model's performance. When $N = 4$ and $T = 100$, the model with the largest degree of shrinkage does best, but using a smaller degree of shrinkage does better as T is raised (for fixed N). The benefit from shrinkage is particularly sizeable when the number of models is large as when $N = 20$. When $N = 20, T = 100$ the reduction in the MSE due to using the strongest degree of shrinkage is close to 10%.

Persistence in the factor dynamics−introduced in the third Monte Carlo experiment−leads to deteriorating performance across all forecasting schemes but means that the relative performance of the simple Granger-Ramanathan combination improves. The methods that constrain the combination weights to sum to unity now produce particularly poor forecasting performance, while the simple projection on the equal-weighted forecast and the shrinkage methods continue to outperform albeit more marginally in the case of the shrinkage schemes.

Introducing heterogeneity in the factor loadings of the various forecasts by drawing these from a beta distribution has two effects. First, it means that the relative performance now differs across forecasting models: The models with larger factor loadings have a higher R^2 than the models with low factor loadings. This means that the approach of choosing the best model now does better (but still underperforms) relative to the results under the unconstrained OLS combination. Secondly, the combination schemes that are based on equal weights now perform worse. This follows from our earlier theoretical results which showed that (generically) equal weights are optimal only when the forecasts errors have identical variances with the same correlations. The effect of weak heterogeneity ($\beta_{if} \sim Beta(5, 5)$) on the performance of the various combination schemes (experiment 4) is quite minor. However, as stronger heterogeneity ($\beta_{if} \sim Beta(1, 1)$) is introduced in the distribution of factor loadings of the underlying forecasting models (experiment 5), clearly equal-weighting progressively performs worse.

When half the forecasts track factor one while the remaining forecasts track factor two (experiment 6), the benefits from combining over using the single best model (which can only track one factor at a time) tend to be particularly large. Moreover, the projection on

equal weights performs particularly well and the shrinkage forecasts also continue to perform well relative to the benchmark.

Finally, when we let half of the forecasts be biased with a bias equal to one-half of the standard deviation terms, the efficiency gain due to omitting an intercept from the combination regression is now more than out-done by the resulting bias. This explains why the general Granger-Ramunathan scheme (GR1) which includes an intercept term now produces better results than the constrained Granger-Ramanathan regressions. Selecting the previous best model is now an even worse idea compared to the results in experiment 2 as there is always the chance of selecting a biased model. Similarly, the equal-weighted scheme produces worse performance than in the case without bias. Since equal weights are the point towards which the least squares estimates are shrunk, this also explains why the three shrinkage schemes perform worse than in the case without bias and now they generally produce worse results than the Granger-Ramanathan benchmark. In contrast the performance of the combination that uses a projection on an intercept and the equal-weighted forecast is unchanged compared with the results for experiment number 2 as only the intercept is changed.

7 Combination of Interval and Probability Distribution Forecasts

So far we have focussed on combining point forecasts. This, of course, reflects the fact that the vast number of academic studies on forecasting only consider point forecasts. However, there has been a growing interest in studying interval and probability distribution forecasts and an emerging literature in economics is considering the scope for using combination methods for such forecasts. This is preceded by the use of combined probability forecasting in areas such as meteorology, c.f. Sanders (1963). Genest and Zidek (1986) present a broad survey of various techniques in this area.

Forecast users are generally interested in the full predictive distribution of the target variable or at least higher order moments or quantiles of this distribution. To see this,

consider the case with lin-lin loss, i.e. $L(e_{t+h,t}) = (\theta 1_{e_{t+h,t} \leq 0} + (1-\theta)1_{e_{t+h,t}>0})|e_{t+h}|$, where $\theta \in [0;1]$. For a given value of θ, the optimal forecast will simply be the θth quantile. Now suppose that the forecast is generated under the knowledge that the forecast user has lin-lin loss but without knowing the particular value of θ. In this situation a forecast must be generated for each feasible value of θ. As θ is varied from zero to one, the entire predictive density is tracked and so the only sufficient statistic in this case is the full predictive density.

To capture the idea that point forecasts generally provide insufficient information for decision makers−other than those with MSE loss−we outline the decision problem underlying a generic forecasting situation. We represent the decision maker's objectives through a utility function $U(Y_{t+h}, \mathbf{d}_t)$ that depends on Y_{t+h} in addition to the forecast user's p-vector of actions, $\mathbf{d}_t = \mathbf{d}(\mathcal{I}_t) \in \mathcal{D}_t$. The objective is to choose actions as a function of current information, $\mathcal{I}_t$, to maximize expected utility:

$$\max_{\mathbf{d}_t \in \mathcal{D}_t} \int U(Y_{t+h}, \mathbf{d}(\mathcal{I}_t)) dF_{yt+h,t}, \tag{84}$$

where $F_{yt+h,t} = \Pr(Y_{t+h} \leq y_{t+h}|\mathcal{I}_t)$ is the conditional distribution function of Y_{t+h} - often taken as reflecting the decision maker's subjective views. This approach allows us to tailor the forecast combination to a specific decision maker's loss and action rule. In the most general case the forecasts being combined comprise a set of N predictive densities, $\{F_{1t+h,t}, F_{2t+h,t}, ..., F_{Nt+h,t}\}$. Let $F^c_{yt+h,t} = f(F_{1t+h,t}, F_{2t+h,t}, ..., F_{Nt+h,t}; \boldsymbol{\omega}_f)$ be the combined density forecast which reflects the individual density forecasts and a set of combination parameters, $\boldsymbol{\omega}_f$. For a given $F^c_{yt+h,t}$, the optimal decision satisfies

$$\int \frac{\partial U(Y_{t+h}, \mathbf{d}(F^c_{yt+h,t}))}{\partial \mathbf{d}} dF^c_{yt+h,t} = 0. \tag{85}$$

The optimal combination of distributions depends on the shape of the utility function, $U(.)$ as well as the form of the decision rule $d(.)$ and its sensitivity to the density forecast $F^c_{yt+h,t}$. Of course $F^c_{yt+h,t}$ cannot be chosen freely since this would lead to unbounded expected utility unless the combination weights are constrained to a compact set. Rather, a good probability forecast combination , $\hat{F}^c_{yt+h,t}$, is one that for all alternative probability forecasts, $\tilde{F}^c_{yt+h,t}$, satisfies

$$\int U(Y_{t+h}, \mathbf{d}(\hat{F}^c_{yt+h,t})) dF_{yt+h,t} \geq \int U(Y_{t+h}, \mathbf{d}(\tilde{F}^c_{yt+h,t})) dF_{yt+h,t}.$$

Notice that expected utility is evaluated under the true probability distribution, F, so a good probability forecast combination is one that leads to good actions.

Define the risk function, $r(F^c_{yt+h,t}, \mathbf{d}(F^c_{yt+h,t}))$, as the negative of the expected utility given $F^c_{yt+h,t}$ and the decision rule, $\mathbf{d}(.)$:

$$r(F^c_{yt+h,t}, \mathbf{d}(F^c_{yt+h,t})) = -\int \partial U(Y_{t+h}, \mathbf{d}(F^c_{yt+h,t}))dF^c_{yt+h,t}.$$

Chamberlain (2000) considers a minimax criterion for the choice of a decision rule $\mathbf{d}(F)$. This requires $\mathbf{d}()$ to be nearly optimal for some set of conditional probability distributions, F, or data generating processes belonging to the set $\mathcal{F}$. A decision rule $\mathbf{d}()$ is said to be $\mathcal{F} - \varepsilon$ risk robust provided that, for some $\varepsilon > 0$,

$$\sup_{F \in \mathcal{F}} \left[r(F, \mathbf{d}(F)) - \inf_{\mathbf{d}_t \in \mathcal{D}_t} r(F, \mathbf{d}(F)) \right] < \varepsilon.$$

Decision rules satisfying this criterion are within ε of the optimal decision rule provided the true conditional density belongs to the set $\mathcal{F}$.

7.1 The Combination Decision

Again it is natural to ask whether the best strategy is to use only a single probability forecast or a combination of these. The notion of forecast encompassing generalizes from point to density forecasts as follows. Suppose we are considering combining N distribution forecasts $f_1, ..., f_N$ whose joint distribution with y is $P(y, f_1, f_2, ..., f_N)$. Factoring this into the product of the conditional distribution of y given $f_1, ..., f_N$, $P(y|f_1, ..., f_N)$, and the marginal distribution of the forecasts, $P(f_1, ..., f_N)$, we have

$$P(y, f_1, f_2, ..., f_N) = P(y|f_1, ..., f_N)P(f_1, ..., f_N). \tag{86}$$

A probability forecast that does not provide information about y given all the other probability density forecasts is referred to as extraneous by Clemen, Murphy and Winkler (1995). If the ith forecast is extraneous we must have

$$P(y|f_1, f_2, ..., f_N) = P(y|f_1, f_2, .., f_{i-1}, f_{i+1}, ..., f_N). \tag{87}$$

If (87) holds, forecast f_i does not contain any information that is useful for forecasting y given the other $N-1$ forecasts. Only if forecast i does not satisfy (87) does it follow that this model is not encompassed by the other models. Interestingly, adding more forecasting models (i.e. increasing N) can lead a previously extraneous model to become non-extraneous if it contains information about the relationship between the existing $N-1$ methods and the new forecast.

For pairwise comparison of probability forecasts, Clemen et al (1995) define the concept of sufficiency. This concept is important because if forecast 1 is sufficient for forecast 2, then its forecasts will be of greater value to all users than forecast 2. Conversely, if neither model is sufficient over the other we would expect some forecast users to prefer model 1 while others prefer model 2.Consider two probability forecasts, $f_1 = P_1(x = 1)$ and $f_2 = P_2(x = 1)$ of some event, X, where $x = 1$ if the event occurs while it is zero otherwise. Also let $v_1(f) = P(f_1 = f)$ and $v_2(g) = P(f_2 = g)$, where $f, g \in \mathcal{F}$, the set of permissible probabilities, $\mathcal{F}$. Forecast 1 is then said to be sufficient for forecast 2 if there exists a stochastic transformation $\zeta(g|f)$ such that for all $g \in \mathcal{F}$,

$$\begin{aligned} \sum_f \zeta(g|f)v_1(f) &= v_2(g), \\ \sum_f \zeta(g|f)fv_1(f) &= gv_2(g). \end{aligned}$$

The function $\zeta(g|f)$ is said to be a stochastic transformation provided that it lies between zero and one and integrates to unity. It represents an additional randomization that has the effect of introducing noise into the first forecast.

7.2 Combination Schemes

Combinations of probability distribution forecasts impose new requirements beyond those we saw for combinations of point forecasts, namely that the combination must be convex with weights confined to the zero-one interval so that the probability forecast never becomes negative and always sums to one.

This still leaves open a wide variety of combination schemes. An obvious way to combine a collection of probability forecasts $\{F_{1,t+h,t}, ..., F_{N,t+h,t}\}$ is through the convex combination

("linear opinion pool"):

$$\bar{F}^c = \sum_{i=1}^{N} \omega_{i,t+h,t} F_{i,t+h,t}, \tag{88}$$

with $0 \leq \omega_{i,t+h,t} \leq 1$ $(i = 1, ..., N)$ and $\sum_{i=1}^{N} \omega_{i,t+h,t} = 1$ to ensure that the combined probability forecast is everywhere non-negative and integrates to one. The generalized linear opinion pool adds an extra probability forecast, $F_{0,t+h,t}$, and takes the form

$$\bar{F}^c = \sum_{i=0}^{N} \omega_{i,t+h,t} F_{i,t+h,t}, \tag{89}$$

where $F_{0,t+h,t}$ can be shown to exist under conditions discussed by Genest and Zidek (1986). Under this scheme the weights are allowed to be negative $\omega_0, \omega_1, ..., \omega_n \in [-1, 1]$ although they still are restricted to sum to unity: $\sum_{i=0}^{N} \omega_{i,t+h,t} = 1$.

Alternatively, one can adopt a logarithmic combination of densities

$$\bar{f}^l = \prod_{i=1}^{N} f_{t+h,t,i}^{\omega_{t+h,t,i}} / \int \prod_{i=1}^{N} f_{t+h,t,i}^{\omega_{t+h,t,i}} d\mu, \tag{90}$$

where $\{\omega_{t+h,t,1}, ..., \omega_{t+h,t,N}\}$ are weights chosen such that the integral is finite and μ is the underlying probability measure. This combination is less dispersed than the linear combination and is also unimodal, c.f. Genest and Zidek (1986).

7.3 Bayesian Methods

Bayesian approaches have been widely used to construct combinations of probability forecasts. Suppose that we have specified a prior, $f(\omega)$ along with a likelihood over the underlying densities, $l(f_1, ..., f_N|\omega)$. From Bayes theorem we then have

$$f(\omega|p_1, ..., p_N) \propto f(\omega) l(f_1, ..., f_N|\omega). \tag{91}$$

Bunn and Mustafaoglu (1978) consider forecasting the risk of political events using a Bayesian approach. Let p_{ik} be the subjective probability of event i by expert k.[8] Suppose that this probability is generated by a beta density reflecting n_{ik} draws, r_{ik} of which are 'successes':

$$f(p_{ik}|r_{ik}, n_{ik}) = p_{ik}^{r_{ik}-1}(1 - p_{ik})^{n_{ik}-r_{ik}-1} \frac{\Gamma(n_{ik})\Gamma(r_{ik})}{\Gamma(n_{ik} - r_{ik})}. \tag{92}$$

[8] Unlike Bunn and Mustafaoglu we do not explicitly condition on multiple information factors.

Hence $\bar{p}_i = r_{ik}/n_{ik}$ reflects the average value of p_{ik} while n_{ik} reflects the precision of the ith probability. From properties of composite beta distributions, $\bar{p}_i$ will be unbiased with mean

$$\bar{p}_i = \frac{\sum_{k=1}^{N} n_{ik}\bar{p}_{ik}}{\sum_{k=1}^{N} n_{ik}}. \tag{93}$$

Min and Zellner (1993) propose combinations based on posterior odds ratios. Let p_1 and p_2 be the posterior probabilities of two models (a fixed parameter and a time-varying parameter model in their application) while $k = p_1/p_2$ is the posterior odds ratio of the two models. Assuming that the two models, M_1 and M_2, are exhaustive the proposed combination scheme has a conditional mean of

$$\begin{aligned} E[Y] &= p_1 E[Y|M_1] + (1-p_1)E[Y|M_2] \\ &= \frac{k}{1+k}E[Y|M_1] + \frac{1}{1+k}E[Y|M_2]. \end{aligned} \tag{94}$$

Palm and Zellner (1992) propose a combination method that accounts for the full correlation structure between the forecast errors. They model the forecast errors from the individual models as follows

$$y_t - \hat{y}_{it} = \theta_i + \varepsilon_{it} + \eta_t, \tag{95}$$

where θ_i is the bias in the ith model's forecast−reflecting perhaps the forecaster's asymmetric loss, c.f. Zellner (1986)− ε_{it} is an idiosyncratic forecast error and η_t is a common component in the forecast errors reflecting an unpredictable component of the outcome variable. It is assumed that both $\varepsilon_{it} \sim N(0, \sigma_i^2)$ and $\eta_t \sim N(0, \sigma_\eta^2)$ are serially uncorrelated (as well as mutually uncorrelated) Gaussian variables with zero mean.

For the case with zero bias ($\theta_i = 0$), Winkler (1981) shows that when $\varepsilon_{it}+\eta_t$ $(i = 1, ..., N)$ has known covariance matrix, $\mathbf{\Sigma}_0$, then the predictive density function of y_t given an N-vector of forecasts $\hat{\mathbf{y}}_t = (\hat{y}_{1t}, ..., \hat{y}_{Nt})'$ is Gaussian with mean $\boldsymbol{\iota}'\mathbf{\Sigma}_0\hat{\mathbf{y}}_t/\boldsymbol{\iota}'\mathbf{\Sigma}_0\boldsymbol{\iota}$ and variance $\boldsymbol{\iota}'\mathbf{\Sigma}_0\boldsymbol{\iota}$. When the covariance matrix of the N time-varying parts of the forecast errors $\varepsilon_{it}+\eta_t$, $\mathbf{\Sigma}$, is unknown but has an inverted Wishart prior $IW(\mathbf{\Sigma}|\mathbf{\Sigma}_0, \delta_0, N)$ with $\delta_0 \geq N$, the predictive distribution of y_{t+1} given $\hat{\mathbf{y}}_{t+1}$ and $\mathcal{I}_t = \{y_1, ..., y_t, \hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_t)$ is a univariate student-t with $\delta_0 + N - 1$ degrees of freedom, mean of $m^* = \boldsymbol{\iota}'\mathbf{\Sigma}_0^{-1}\hat{\mathbf{y}}_t/\boldsymbol{\iota}'\mathbf{\Sigma}_0^{-1}\boldsymbol{\iota}$ and variance $(\delta_0+N-1)s^{*2}/(\delta_0+N-3)$, where $s^{*2} = (\delta_0 + (m^*\boldsymbol{\iota} - \hat{\mathbf{y}}_t)'\mathbf{\Sigma}_0^{-1}(m^*\boldsymbol{\iota} - \hat{\mathbf{y}}_t))/(\delta_0 + N - 1)\boldsymbol{\iota}'\mathbf{\Sigma}_0^{-1}\boldsymbol{\iota}$.

Palm and Zellner (1992) extend these results to allow for a non-zero bias. The structure of the forecast errors (95) is reflected in a Wishart prior for $\mathbf{\Sigma}^{-1}$ with v degrees of freedom and covariance matrix $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_{\varepsilon 0} + \sigma^2_{\eta 0}\boldsymbol{\iota}\boldsymbol{\iota}'$ (with known parameters $\mathbf{\Sigma}_{\varepsilon 0}, \sigma^2_{\eta 0}$):

$$P(\mathbf{\Sigma}^{-1}) \propto |\mathbf{\Sigma}^{-1}|^{(v-N-1)/2}|\mathbf{\Sigma}_0^{-1}|^{-v/2}\exp(-\frac{1}{2}tr(\mathbf{\Sigma}_0\mathbf{\Sigma}^{-1})).$$

Assuming a likelihood function

$$L(\boldsymbol{\theta}, \mathbf{\Sigma}^{-1}|\mathcal{I}_t) =\propto |\mathbf{\Sigma}^{-1}|^{-T/2}\exp(-\frac{1}{2}tr(\mathbf{S}\mathbf{\Sigma}^{-1}) - \frac{1}{2}tr((\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\boldsymbol{\iota}'\boldsymbol{\iota}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{\Sigma}^{-1})),$$

where $\hat{\boldsymbol{\theta}} = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\mathbf{y}$, $\mathbf{y} =(y_1, ..., y_T)$ and $\mathbf{S} =(\mathbf{y}-\boldsymbol{\iota}\hat{\boldsymbol{\theta}}')'(\mathbf{y}-\boldsymbol{\iota}\hat{\boldsymbol{\theta}}')$, we get the predictive probability distribution function of y_{T+1} given $\hat{\mathbf{y}}_{T+1}, \mathcal{I}_T$:

$$P(y_{T+1}|\hat{\mathbf{y}}_t, \mathcal{I}_t) \propto \left[1 + (y_{T+1}-\mu^{**})^2/(T-1)s^{**2}\right]^{-(T+v)/2},$$

where $\mu^{**} = \boldsymbol{\iota}'\bar{\mathbf{S}}^{-1}\boldsymbol{\mu}/\boldsymbol{\iota}'\bar{\mathbf{S}}^{-1}\boldsymbol{\iota}$, $s^{**2} = 1/(T(T-1))\left[T+1+T(\mu^{**}\boldsymbol{\iota}-\boldsymbol{\mu})'\bar{\mathbf{S}}^{-1}(\mu^{**}\boldsymbol{\iota}-\boldsymbol{\mu})\right]/(\boldsymbol{\iota}'\bar{\mathbf{S}}^{-1}\boldsymbol{\iota})$ and $\bar{\mathbf{S}} = \mathbf{S} + \mathbf{\Sigma}_0$.

7.3.1 Bayesian Model Averaging

Bayesian Model Averaging methods have been proposed by, inter alia, Leamer (1978), Draper (1995) and Hoeting et al. (1999) and are increasingly used in empirical studies, see e.g. Jacobson and Karlsson (2003). Under this approach, the predictive density can be computed by averaging over a set of models, $i = 1, ..., N$, each characterized by parameters $\boldsymbol{\theta}_i$:

$$f\left(y_{t+h}\left|\mathcal{I}_t\right.\right) = \sum_{i=1,..,N} \Pr\left(M_i\left|\mathcal{I}_t\right.\right) f_i\left(y_{t+h}, \boldsymbol{\theta}_i\left|\mathcal{I}_t\right.\right), \tag{96}$$

where $\Pr\left(M_i\left|\mathcal{I}_t\right.\right)$ is the posterior probability of model M_i obtained from the prior distributions, the model priors $\Pr\left(M_i\right)$, the priors for the unknown parameters, $\Pr\left(\boldsymbol{\theta}_i\left|M_i\right.\right)$, and the likelihood functions of the models under consideration. $f_i\left(y_{t+h}, \boldsymbol{\theta}_i\left|\mathcal{I}_t\right.\right)$ is the predictive density of y_{t+h} and $\boldsymbol{\theta}_i$ under the ith model, given information at time t, $\mathcal{I}_t$.

7.4 Combinations of Interval Forecasts

Combinations of interval forecasts do not raise any new issues and can effectively be viewed as combining two point forecasts, namely the lower and upper bound for the confidence interval.

Suppose that we have N interval forecasts each taking the form of a lower and an upper limit $\{l_{t+h,t,i}; u_{t+h,t,i}\}$. A combined interval forecast can then be computed as $\{\bar{l}^c_{t+h,t,i}; \bar{u}^c_{t+h,t,i}\}$, where

$$\begin{aligned} \bar{l}^c_{t+h,t,i} &= \sum_{i=1}^{N} \omega^l_{t+h,t,i} l_{t+h,t,i}, \\ \bar{u}^c_{t+h,t,i} &= \sum_{i=1}^{N} \omega^u_{t+h,t,i} u_{t+h,t,i}. \end{aligned} \tag{97}$$

Although it is not required that the combination weights be confined to $[0; 1]$ and sum to unity, this is a natural constraint to impose since each interval represents a coverage probability. This interpretation of the combined interval need not be preserved unless the weights sum to one. In principle we could use different weights on the lower and upper limits, such as when a model generates a precise estimate of the lower limit, but not of the upper limit. Whether this will work in practice is less clear, however, since one would expect that the limits associated with a particular model are connected and certainly not unique unless additional constraints (e.g. that each interval forecast minimizes the length of the interval) are imposed.

8 Conclusion

In his classical survey of forecast combinations, Clemen (1989, p. 567) concluded that "Combining forecasts has been shown to be practical, economical and useful. Underlying theory has been developed, and many empirical tests have demonstrated the value of composite forecasting. We no longer need to justify this methodology."

In the early days of the combination literature the set of forecasts was taken as given, but recent experiments undertaken by Stock and Watson (2001, 2003) and Marcellino (2004) let the forecast user control both the number of forecasting models as well as the types of forecasts that are being combined. This opens a whole new set of issues: is it best to combine linear models with different regressors or is it better to combine different families of forecasting models, e.g. linear and nonlinear, or maybe the same model using estimators with varying degrees of robustness? The answer to this depends of course on the type of

misspecification the model combination can hedge against. Unfortunately this is typically unknown so general answers are hard to come by.

Since then, combination methods have gained even more ground in the forecasting literature, largely because of the strength of the empirical evidence suggesting that these methods systematically perform better than alternatives based on forecasts from a single model. Stable, equal weights have so far been the workhorse of the combination literature and have set a benchmark that has proved surprisingly difficult to beat. This is surprising since−on theoretical grounds−one would not expect any particular combination scheme to be dominant, since the various methods incorporate restrictions on the covariance matrix that are designed to trade off bias against reduced parameter estimation error. The optimal bias can be expected to vary across applications, and the scheme that provides the best trade-off is expected to depend on the sample size, the number of forecasting models involved, the ratio of the variance of individual models' forecast errors as well as their correlations and the degree of instability in the underlying data generating process.

Current research also provides encouraging pointers towards modifications of this simple strategy that can improve forecasting. Modest time-variations in the combination weights and trimming of the worst models have generally been found to work well, as has shrinkage towards equal weights or some other target requiring the estimation of only few parameters, particularly in applications with combinations of large numbers of forecasts.

References

[1] Aiolfi, M. and C. A. Favero, 2003, Model Uncertainty, Thick Modeling and the Predictability of Stock Returns. Forthcoming in Journal of Forecasting.

[2] Aiolfi, M. and A. Timmermann, 2004, Persistence of Forecasting Performance and Combination Strategies. Mimeo, UCSD.

[3] Armstrong, J.S., 1989, Combining Forecasts: The End of the Beginning or the Beginning of the End, International Journal of Forecasting, 5, 585-588.

[4] Bates, J.M. and C.W.J. Granger, 1969, The Combination of Forecasts. Operations Research Quarterly 20, 451-468.

[5] Bunn, D.W., 1975, A Bayesian Approach to the Linear Combination of Forecasts, Operations Research Quarterly, 26, 325-29.

[6] Bunn, D.W., 1985, Statistical Efficiency in the Linear Combination of Forecasts, International Journal of Forecasting, 1, 151-163.

[7] Bunn, D.W. and M.M. Mustafaoglu, 1978, Forecasting Political Risks. Management Science 24, 1557-1567.

[8] Chamberlain, G., 2000, Econometrics and Decision Theory. Journal of Econometrics 95, 255-283.

[9] Chan, Y.L, J.H. Stock and M.W. Watson, 1999, A Dynamic factor model framework for forecast combination. Spanish Economic Review 1, 91-122.

[10] Chong, Y.Y. and D.F. Hendry, 1986, "Econometric Evaluation of Linear Macro-Economic Models," Review of Economic Studies, 53, 671-690.

[11] Christoffersen, P. and F.X. Diebold, 1997, Optimal Prediction under Asymmetrical Loss. Econometric Theory 13, 806-817.

[12] Clemen, R.T., 1987, Combining Overlapping Information. Management Science 33, 3, 373-380.

[13] Clemen, R.T., 1989, Combining Forecasts: A Review and Annotated Bibliography. International Journal of Forecasting 5, 559-581.

[14] Clemen, R.T. and R.L. Winkler, 1986, Combining Economic Forecasts, Journal of Business and Economic Statistics, 4, 39-46.

[15] Deutsch, M., C.W.J. Granger and T. Terasvirta, 1994. The Combination of Forecasts using Changing Weights. International Journal of Forecasting 10, 47-57.

[16] Diebold, F.X., 1988, Serial Correlation and the Combination of Forecasts. Journal of Business and Economic Statistics 6, 105-111.

[17] Diebold, F.X., 1989, Forecast Combination and Encompassing: Reconciling Two Divergent Literatures, International Journal of Forecasting, 5, 589-92.

[18] Diebold, F. X. and J. A. Lopez, 1996, Forecast Evaluation and Combination. In Maddala and Rao (eds.) Handbook of Statistics. Elsevier: Amsterdam.

[19] Diebold, F.X. and P. Pauly, 1987, Structural Change and the Combination of Forecasts. Journal of Forecasting 6, 21-40.

[20] Diebold, F.X. and P. Pauly, 1990, The Use of Prior Information in Forecast Combination, International Journal of Forecasting, 6, 503-508.

[21] Donaldson, R.G. and M. Kamstra, 1996, Forecast Combining with Neural Networks. Journal of Forecasting 15, 49-61.

[22] Dunis, C. J. Laws and S. Chauvin, 2001, The Use of Market Data and Model Combinations to Improve Forecast Accuracy. Page 45-80 in Dunis, Timmermann and Moody (eds) (2001).

[23] Dunis, C.L., A. Timmermann, and J.E. Moody. (eds), 2001, Developments in Forecasts Combination and Portfolio Choice. Oxford: Wiley.

[24] Elliott, G., 2004, Forecast Combination with Many Forecasts. Mimeo, UCSD.

[25] Elliott, G. and A. Timmermann, 2003, Optimal Forecast Combination Weights Under Regime Switching. Forthcoming, International Economic Review.

[26] Elliott, G. and A. Timmermann, 2004, Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions. Journal of Econometrics 122, 47-79.

[27] Figlewski, S. and T. Urich, 1983, "Optimal Aggregation of Money Supply Forecasts: Accuracy, Profitability and Market Efficiency," Journal of Finance, 28, 695-210.

[28] Genest, S. and J. Zidek, 1986, Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science 1, 114-148.

[29] Giacomini, R. and I. Komunjer, 2002, Evaluation and Combination of Conditional Quantile Forecasts, UCSD working paper.

[30] Granger, C.W.J., 1989, Combining Forecasts - Twenty Years Later. Journal of Forecasting 8, 167-173.

[31] Granger, C.W.J., and M. Machina, 2004,

[32] Granger, C.W.J. and M.H. Pesaran, 2000, Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting 19, 537-560.

[33] Granger, C.W.J. and R. Ramanathan, 1984, Improved Methods of Combining Forecasts. Journal of Forecasting 3, 197-204.

[34] Granger, C.W.J. and Y. Jeon, 2004, Thick Modeling, Economic Modelling, 21, 323-343.

[35] Guidolin, M. and A. Timmermann, 2003, An Econometric Model of Nonlinearities in Bond and Stock Prices. Forthcoming in Journal of Applied Econometrics.

[36] Gupta, S. and P.C. Wilton, 1987, Combination of Forecasts: An Extension, Management Science, 33, 356-372.

[37] Hendry, D.F. and M.P. Clements, 2002, Pooling of Forecasts. Econometrics Journal 5, 1-26.

[38] Hoeting, J. A., D. Madigan, A.E. Raftery and C.T. Volinsky, 1999, Bayesian Model Averaging: A Tutorial. Statistical Science, 14, 382-417.

[39] Jaganathan, R. and T. Ma, 2003, Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps, Journal of Finance 1651-1684.

[40] Jacobson, T. and S. Karlsson, 2003, "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," Forthcoming, Journal of Forecasting.

[41] Jobson, J.D. and B. Korkie, 1980, Estimation for Markowitz Efficient Portfolios. Journal of American Statistical Association 75 (371), 544-554

[42] Kang, H., 1986, Unstable Weights in the Combination of Forecasts. Management Science 32, 683-695.

[43] Leamer, E., 1978, Specification Searches. Wiley.

[44] Ledoit, O. and M. Wolf, 2003, Improved Estimation of the Covariance Matrix of stock Returns with an Application to Portfolio Selection. Journal of Empirical Finance 10, 603-621.

[45] Ledoit, O. and M. Wolf, 2004, Honey, I shrunk the Sample Covariance Matrix. Forthcoming Journal of Portfolio Management.

[46] LeSage, J.P., and M. Magura, 1992, "A Mixture-Model Approach to Combining Forecasts," Journal of Business and Economic Statistics, 10, 445-453.

[47] Makridakis, S., 1989, "Why Combining Works?," International Journal of Forecasting, 5, 601-603.

[48] Makridakis, S. and M. Hibon, 2000, "The M3-Competition: Results, Conclusions and Implications," International Journal of Forecasting, 16, 451-476.

[49] Makridakis, S. and R.L. Winkler, 1983, Averages of Forecasts: Some Empirical Results. Management Science 29, 987-996.

[50] Marcellino, M., 2004, Forecast Pooling for Short Time Series of Macroeconomic Variables. Oxford Bulletin of Economic and Statistics, 66, 91-112.

[51] McNees, S.K., 1992, "The Uses and Abuses of "Consensus" Forecasts," Journal of Forecasting, 11, 703-710.

[52] Min, C-k, and A. Zellner, 1993, Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates. Journal of Econometrics 56, 89-118.

[53] Newbold, P. and D.I. Harvey, 2001, Forecast Combination and Encompassing. In Clements, M.P. and D.F. Hendry (eds), A Companion to Economic Forecasting. Oxford: Blackwells.

[54] Palm, F. C. and A. Zellner, 1992, To Combine or not to Combine? Issues of Combining Forecasts. Journal of Forecasting 11, 687-701.

[55] Raftery, A.E., D. Madigan and J.A. Hoeting, 1997, Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association 92, 179-191.

[56] Reid, D.J., 1968, Combining three estimates of Gross Domestic Product. Economica 35, 431-444.

[57] Sanders, F., 1963, On Subjective probability forecasting. Journal of Applied Meteorology 2, 196-201.

[58] Sessions, D.N. and S.Chattererjee, 1989, "The Combining of Forecasts Using Recursive Techniques with Nonstationary Weights," Journal of Forecasting, 8, 239-251.

[59] Stock, J.H. and M. Watson, 2001, A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. Pages 1-44 In R.F. Engle and H. White (eds). Festschrift in Honour of Clive Granger.

[60] Stock, J.H. and M. Watson, 2003, Combination Forecasts of Output Growth in a Seven-Country Data Set. Forthcoming in Journal of Forecasting.

[61] Swanson, N.R., and T. Zeng, 2001, Choosing among Competing Econometric Forecasts: Regression-Based Forecast Combination Using Model Selection, Journal of Forecasting, 6, 425-440.

[62] Winkler R.L., 1981, Combining probability distributions from dependent information sources. Management Science 27, 479-488.

[63] Winkler, R.L., 1989, Combining Forecasts: A Philosophical Basis and Some Current Issues, International Journal of Forecasting, 5, 605-609.

[64] Winkler, R.L. and R.T. Clemen, 1992, Sensitivity of Weights in Combining Forecasts, Operations Research, 40(3), 1992.

[65] Winkler, R.L. and S. Makridakis, 1983, The Combination of Forecasts, Journal of the Royal Statistical Society Series A, 146, 150-57.

[66] Wright, S.M, and S.E. Satchell, 2003, Generalized mean-variance analysis and robust portfolio diversification. Pages 40-54 in S.E Satchell and A. Scowcroft (eds.) Advances in portfolio construction and implementation. Butterworth Heinemann, London.

[67] Yang, Y., 2004, Combining Forecasts Procedures: Some Theoretical Results, Econometric Theory, 20, 176-190.

[68] Zellner, A., 1986, Biased Predictors, rationality and the evaluation of forecasts. Economics Letters 21, 45-48.

[69] Zellner, A., C. Hong and C-k Min, 1991, Forecasting Turning Points in International Output Growth Rates using Bayesian Exponentially Weighted Autoregression, Time-varying Parameter, and Pooling Techniques. Journal of Econometrics 49, 275-304.

Table 2. Simulation results from forecast combinations under factor structure

number of models	sample size	GR1	GR2	GR3	previous best	simple EW	Projection EW	Shrink 1	Shrink 2	Shrink 3
Experiment 1: Equal weights summing to one										
4	100	1	0.991	0.981	1.162	0.954	0.972	0.990	0.989	0.988
4	500	1	0.998	0.996	1.237	0.990	0.994	0.998	0.998	0.998
4	1000	1	0.999	0.997	1.268	0.995	0.999	0.999	0.999	0.999
10	100	1	0.991	0.982	1.180	0.895	0.910	0.986	0.981	0.971
10	500	1	0.997	0.995	1.358	0.978	0.983	0.997	0.997	0.997
10	1000	1	0.999	0.998	1.389	0.993	0.995	0.999	0.999	0.999
20	100	1	0.992	0.981	1.123	0.805	0.819	0.968	0.946	0.907
20	500	1	0.997	0.994	1.373	0.955	0.961	0.996	0.995	0.993
20	1000	1	0.999	0.998	1.385	0.979	0.981	0.999	0.998	0.998
Experiment 2: Equal weights										
4	100	1	0.991	1.138	1.234	1.107	0.972	0.990	0.990	0.988
4	500	1	0.998	1.148	1.344	1.142	0.995	0.998	0.998	0.998
4	1000	1	0.999	1.165	1.363	1.164	1.000	0.999	0.999	0.999
10	100	1	0.989	1.267	1.296	1.151	0.918	0.984	0.979	0.972
10	500	1	0.998	1.304	1.557	1.283	0.985	0.998	0.998	0.997
10	1000	1	0.999	1.320	1.628	1.310	0.994	0.999	0.999	0.999
20	100	1	0.990	1.349	1.192	1.100	0.809	0.967	0.947	0.920
20	500	1	0.997	1.380	1.606	1.330	0.962	0.997	0.996	0.995
20	1000	1	0.998	1.374	1.659	1.346	0.980	0.998	0.997	0.997
Experiment 3: Factor dynamics										
4	100	1	0.984	1.358	1.680	1.315	0.969	0.984	0.983	0.982
4	500	1	0.998	1.353	1.859	1.345	0.995	0.998	0.998	0.998
4	1000	1	0.998	1.386	1.928	1.384	0.999	0.998	0.998	0.998
10	100	1	0.993	1.946	2.352	1.766	0.919	0.989	0.986	0.985
10	500	1	0.997	1.975	2.863	1.940	0.983	0.996	0.996	0.996
10	1000	1	0.999	2.065	3.037	2.051	0.996	0.999	0.999	0.999
20	100	1	0.990	2.400	2.603	1.966	0.816	0.970	0.960	0.974
20	500	1	0.997	2.479	3.575	2.399	0.960	0.996	0.996	0.996
20	1000	1	0.999	2.440	3.800	2.404	0.981	0.998	0.998	0.998

Experiment 4: Weak heterogeneity

4	100	1	0.992	1.106	1.180	1.120	0.980	0.991	0.991	0.989
4	500	1	0.998	1.123	1.280	1.160	1.002	0.998	0.998	0.998
4	1000	1	0.999	1.134	1.288	1.179	1.008	0.999	0.999	0.999
10	100	1	0.989	1.164	1.217	1.169	0.928	0.984	0.979	0.972
10	500	1	0.998	1.183	1.389	1.299	0.993	0.998	0.998	0.997
10	1000	1	0.999	1.207	1.439	1.324	1.000	0.999	0.999	0.999
20	100	1	0.990	1.148	1.093	1.107	0.812	0.966	0.946	0.920
20	500	1	0.997	1.178	1.401	1.336	0.964	0.997	0.996	0.995
20	1000	1	0.998	1.172	1.444	1.359	0.987	0.998	0.997	0.997

Experiment 5: Strong heterogeneity

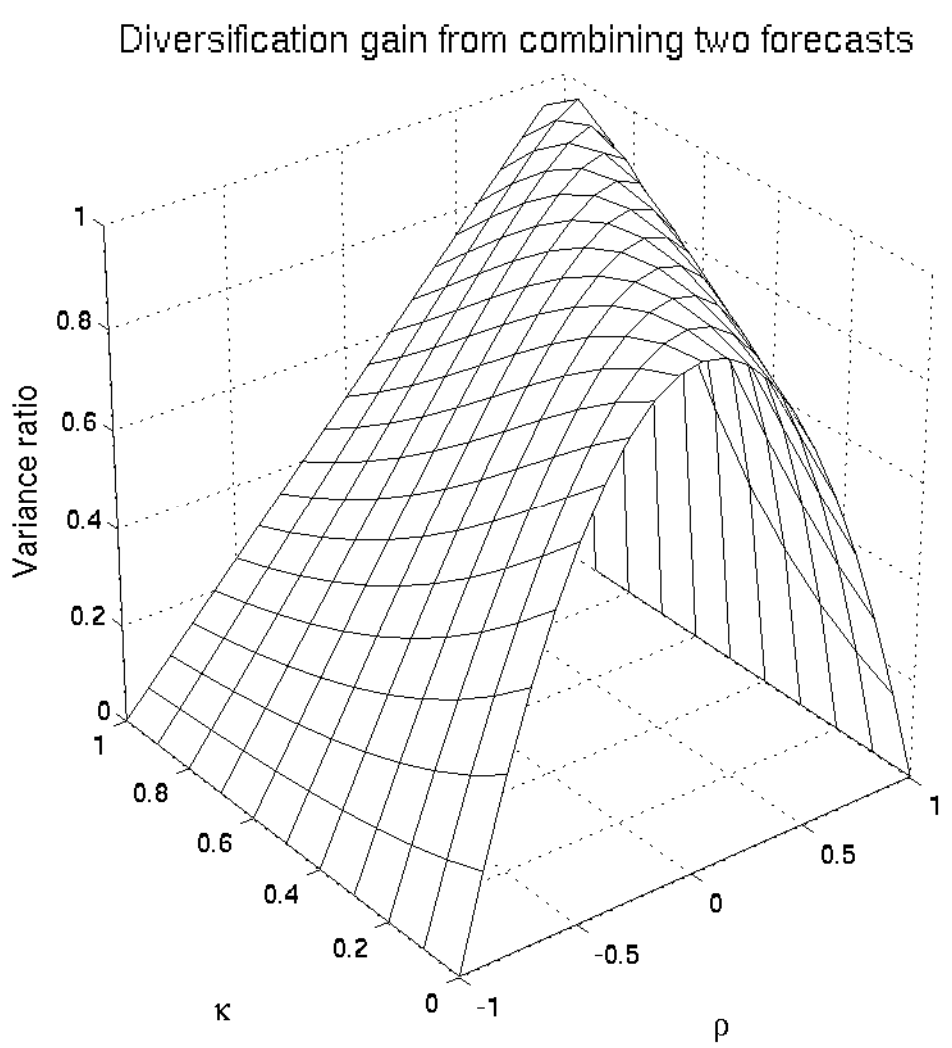
4	100	1	0.992	1.063	1.115	1.178	1.012	0.991	0.990	0.989
4	500	1	0.999	1.076	1.181	1.223	1.037	0.999	0.999	0.999
4	1000	1	0.999	1.081	1.185	1.227	1.029	0.999	0.999	0.999
10	100	1	0.989	1.038	1.076	1.216	0.952	0.984	0.979	0.972
10	500	1	0.998	1.066	1.206	1.353	1.017	0.998	0.998	0.997
10	1000	1	0.999	1.062	1.216	1.384	1.037	0.999	0.999	0.999
20	100	1	0.989	1.009	0.939	1.140	0.826	0.965	0.946	0.921
20	500	1	0.997	1.036	1.167	1.382	0.985	0.997	0.996	0.995
20	1000	1	0.998	1.029	1.179	1.401	1.010	0.998	0.997	0.997

Experiment 6: Block-diagonal factor structure

4	100	1	0.991	1.287	1.320	1.252	0.974	0.990	0.989	0.988
4	500	1	0.998	1.295	1.423	1.288	0.995	0.998	0.998	0.998
4	1000	1	0.999	1.325	1.474	1.321	0.999	0.999	0.999	0.999
10	100	1	0.990	1.552	1.471	1.401	0.913	0.984	0.980	0.974
10	500	1	0.998	1.610	1.735	1.581	0.984	0.998	0.997	0.997
10	1000	1	0.999	1.627	1.823	1.611	0.993	0.999	0.999	0.999
20	100	1	0.991	1.749	1.445	1.427	0.809	0.968	0.952	0.940
20	500	1	0.997	1.791	1.901	1.722	0.960	0.997	0.996	0.995
20	1000	1	0.998	1.783	1.928	1.744	0.981	0.998	0.998	0.997

Experiment 7: Bias in individual forecasts

4	100	1	1.063	1.166	1.274	1.155	0.972	1.062	1.061	1.060
4	500	1	1.057	1.168	1.367	1.177	0.995	1.056	1.056	1.056
4	1000	1	1.077	1.200	1.407	1.209	1.000	1.077	1.077	1.077
10	100	1	1.068	1.294	1.354	1.199	0.918	1.062	1.057	1.048
10	500	1	1.071	1.330	1.575	1.336	0.985	1.071	1.071	1.070
10	1000	1	1.069	1.342	1.639	1.359	0.994	1.069	1.069	1.069
20	100	1	1.034	1.358	1.241	1.139	0.809	1.010	0.989	0.961
20	500	1	1.059	1.401	1.647	1.385	0.962	1.058	1.057	1.056
20	1000	1	1.061	1.394	1.704	1.406	0.980	1.060	1.060	1.060


Diversification gain from combining two forecasts
Variance ratio
1
0.8
0.6
0.4
0.2
0
κ
1
0.8
0.6
0.4
0.2
0
ρ
-1
-0.5
0
0.5
1