Ensembles and combinations: using multiple models to improve forecasts

An increasing size of the toolbox of forecasting methods is available for decision makers. These methods, including statistical and econometric models, machine learning algorithms, and even judgemental forecasting (see an encyclopedic overview by Petropoulos et al., 2020), have their own specialities and are developed under different model specifications with assumptions on the Data Generation Process (DGP) or the associated error distributions. Given a pool of forecasting methods, how to best exploit information in the individual forecasts obtained from these methods?

Several studies in the forecasting literature are devoted to identifying a single 'best model' for a given time series. Given a family of models, information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978), are commonly used for model selection (e.g., Qi and Zhang, 2001; Billah et al., 2005; Yang, 2005). More generally, cross-validation, in its various forms, such as the hold-out approach and the out-of-sample rolling scheme, has been used successfully to select the best forecast when multiple model families or model-free forecasts are considered (e.g., Kohavi et al., 1995; Poler and Mula, 2011; Fildes and Petropoulos, 2015; Inoue et al., 2017; Talagala et al., 2018). However, different criteria may lead to different results of forecast selections. Kourentzes et al. (2019) argued that model selection was a challenging task for two reasons: the sample, parameter and model uncertainty associated with identifying a single best forecast, and the ill-defined best forecast.

(Modified to use multiple models instead of combination ↓.)

Given these challenges, alternatively Bates and Granger (1969) have suggested combining multiple forecasts. The idea of combining forecasts is derived from the simple portfolio diversification argument (Timmermann, 2006), which is a risk management strategy with an obvious intuition: do not put all eggs into one basket. Even though slightly earlier articles have provided empirical justification of the superiority of forecast combinations over individual forecasts (e.g., Barnard, 1963; Crane and Crotty, 1967), the work by Bates and Granger (1969) is often considered to be the seminal article on forecast combinations as they developed a general analysis and further explored more possibilities for forecast combinations by extending a simple average to a weighted combination. Furthermore, the idea of combining forecasts is also widely used in machine learning, referred to as forecast ensembles. Similar to combination, the ensemble is a machine learning

paradigm using multiple models to solve the same problem. It is difficult to trace the beginning of the history of ensemble forecasting. However, it is clear that ensemble techniques have become a hot topic in various fields, especially weather forecasting (see an overview by Leutbecher and Palmer, 2008), since the 1990s. Lewis (2005) provided a genealogy to depict the scientific roots of ensemble forecasting from several fundamental lines of research.

There are nearly five decades of empirical and theoretical investigations support that combining multiple forecasts often achieves improved forecasting performance on average than selecting a single individual forecast. Important early contributions in this area were summarized by Granger (1989), Clemen (1989), Palm and Zellner (1992), and Timmermann (2006). Clemen (1989) surveyed over two hundred statistical literature on forecast combinations and provided a primary conclusion that forecasting accuracy could be substantially improved by combining multiple forecasts. Timmermann (2006) attributed the superiority of forecast combinations over a single model to the fact that individual forecasts obtained based on heterogeneous information sets, may be very differently affected by structural breaks and subject to misspecification bias of unknown form. They further concluded that forecast combinations were beneficial due to diversification gains. More recently, Atiya (2020) illustrated graphically why forecast combinations were superior.

1 Different ways of using multiple models

- Combinations: A (usually linear) combination of forecasts from multiple models are used for one series. This includes combining point forecasts, quantile forecasts or full distributional forecasts. It covers simple averaging, weighted averaging, and sometimes combinations based on ML algorithms. e.g., FFORMA and related methods.

- Ensembles: Although "ensembles" has been used in different ways in different literatures, we will use "ensemble" to mean a mixture of the forecast distributions from multiple models. In many ways this is simpler than combinations as the relationship between the methods can be ignored. Need to discuss when they are equivalent.

- Boosting: Multiple models used for one series in sequence. Equivalent to hybrid forecasting where residuals from one method are modelled using a different method.

- Bagging: One or more models applied to multiple similar series, and then a combination or en-

semble is taken. Bagging requires a method for generating multiple series. Some possibilities are STL-ETS and GRATIS.

- Stacking.

Simple example to illustrate differences. Suppose we have one series and two methods: an ARIMA model and a CNN.

- A combination would apply both to the same series and average the results. Unless we are only interested in point forecasting, the averaging would need to take account of the correlation between the forecast errors.
- An ensemble would apply both to the same series and generate forecast distributions from each. These would then be mixed (possibly with weighting) to form the final forecast distribution.
- Boosting would apply the ARIMA model to the series, and then apply the CNN to the residuals. The final forecasts would be the forecasts from the ARIMA model plus the forecasts from the CNN.
- Bagging would generate multiple series like the series of interest, and apply one of the methods to all the generated series. These could then be combined, or ensembled.

2 Point forecast combinations

2.1 Simple combinations

Considerable literature has accumulated over the years regarding the way in which individual forecasts are combined. A unanimous conclusion is that simple combination schemes are hard to beat (Clemen, 1989; Stock and Watson, 2004; Lichtendahl and Winkler, 2020). More specifically, simple combination rules which ignore past information regarding the precision of individual forecasts and correlations between forecast errors work reasonably well relative to more sophisticated combination schemes, as noted in Clemen's (1989) survey. Lichtendahl and Winkler (2020) attributed this phenomenon to a lower risk of simple combination methods resulting in bad forecasts than more refined combination methods. Timmermann (2006) concisely summarized the reasons for the

success of simple combinations by the importance of parameter estimation error—simple combination schemes did not require estimating parameters such as combination weights based on forecast errors, thus avoiding parameter estimation error that often existed in weighted combinations.

The vast majority of studies on combining multiple models has dealt with point forecasting, even though point forecasts generally provide insufficient information for decision making. The simple average of forecasts based on equal weights stands out as the most popular and surprisingly robust combination rule (see Bunn, 1985; Clemen and Winkler, 1986; Stock and Watson, 2003; Genre et al., 2013). Makridakis et al. (1982) reported the results of M-competition, a forecasting competition involving 1001 economic time series, and found that the simple average outperformed the individual techniques. Clemen (1989) provided an extensive bibliographical review of the early work on the combination of forecasts, and then addressed the issue that the arithmetic means often dominated more refined forecast combinations. Makridakis and Winkler (1983) concluded empirically that the accuracy of combined forecasts was improved and the variability associated with the choice of methods was reduced, as the number of individual methods included in a simple average increased. Palm and Zellner (1992) concisely summarized the advantages of adopting a simple average into three points: (i) combination weights were equal and did not have to be estimated, (ii) a simple average significantly reduced variance and bias by averaging out individual bias in many cases, and (iii) a simple average should be considered when the uncertainty of weight estimation was taken into account. Furthermore, Timmermann (2006) pointed out that the good average performance of the simple average depended strongly on model instability and the ratio of forecast error variances associated with different forecasting models.

More attention has been given to other options, including the median and mode, as well as trimmed means (e.g., Chan et al., 1999; Stock and Watson, 2004; Genre et al., 2013; Jose et al., 2014; Grushka-Cockayne et al., 2017), due to their robustness in the sense of being less affected by extreme forecasts than a simple average (Lichtendahl and Winkler, 2020). There is little consensus in the literature as to whether the mean or the median of individual forecasts performs better in terms of point forecasting (Kolassa, 2011). Specifically, McNees (1992) found no significant difference between the mean and the median, while the results of Stock and Watson (2004) supported the mean and Agnew (1985) recommended the median. Jose and Winkler (2008) studied the forecasting performance of the mean and median, as well as the trimmed and winsorized means. Their results suggested that the trimmed and winsorized means were appealing because of their simplicity and

robust performance. Kourentzes et al. (2014) compared empirically the mean, mode and median combination operators based on kernel density estimation, and found that the three operators dealt with outlying extreme values differently, with the mean being the most sensitive and the mode operator the least. Based on these experimental results, they recommended further investigation of the use of the mode and median operators, which had been largely overlooked in relevant literature.

Compared to various refined combination approaches and advanced machine learning algorithms, simple combinations seem to be outdated and uncompetitive in the big data era. However, the results from the recent M4 competition (Makridakis et al., 2020) showed that simple combinations could achieve fairly good forecasting performance and still be competitive. Specifically, a simple equal-weights combination achieved the third best performance for yearly time series (Shaub, 2019) and a median combination of four models achieved sixth place for the point forecasts (Petropoulos and Svetunkov, 2020). Genre et al. (2013) encompassed a variety of combination methods in the case of forecasting GDP growth and the unemployment rate. They found that the simple average set a high benchmark, with few of the combination schemes outperforming it. Therefore, simple combination rules have been consistently the choice of many researchers and provide a tough benchmark to measure the effectiveness of the newly proposed weight estimation algorithms (e.g., Makridakis and Hibon, 2000; Stock and Watson, 2004; Makridakis et al., 2020; Montero-Manso et al., 2020; Kang et al., 2020; Wang et al., 2021).

2.2 Combination weighting schemes

Though the combined forecasts formed by simple combination rules are acceptable for illustrative and concise purposes, the accumulated evidence of the forecasting literature suggests assigning greater weights to the individual forecasts which contain lower errors. The issue to be addressed is how to best weight the different forecasts used for combination. The general point forecast combination problem can be defined as seeking a one-dimensional aggregator that reduces the information up to time t in an N-vector of h-step-ahead forecasts, $\hat{\mathbf{y}}_{t+h|t} = \left(\hat{y}_{t+h|t,1}, \hat{y}_{t+h|t,2}, \ldots, \hat{y}_{t+h|t,N}\right)'$, to a single combined h-step-ahead forecast $\tilde{y}_{t+h|t} = C\left(\hat{\mathbf{y}}_{t+h|t}; \boldsymbol{w}_{t+h|t}\right)$, where $\boldsymbol{w}_{t+h|t}$ is an N-vector of combination weights. The general class of combination methods represented by the mapping, C, from $\hat{\mathbf{y}}_{t+h|t}$ to y_{t+h}, comprises linear, non-linear, and time-varying combinations. Below we discuss in detail the use of various weighting schemes to determine combination weights.

2.2.1 Linear combinations

Typically, the combined forecast is commonly constructed as a linear combination of the individual forecasts. To this end a combined forecast of the linear form can be written as

$$\tilde{y}_{t+h|t} = \boldsymbol{w}'_{t+h|t}\hat{\mathbf{y}}_{t+h|t}, \tag{1}$$

where $\boldsymbol{w}_{t+h|t} = \left(w_{t+h|t,1}, \ldots, w_{t+h|t,N}\right)'$ is an N-vector of linear combination weights assigned to N individual forecasts.

Optimal weights

The seminal work of Bates and Granger (1969) proposed a method to find 'optimal' weights by minimizing the variance of the combined forecast error, and discussed only the combination of pairs of forecasts. Newbold and Granger (1974) then extended the method to the combination of several forecasts. Specifically, assuming that individual forecasts are unbiased and their variance of errors is consistent over time, the combined forecast obtained by a linear combination will also be unbiased. Differentiating with respect to $\boldsymbol{w}_{t+h|t}$ and solving the first order condition, the variance of the combined forecast error is minimized by taking

$$\boldsymbol{w}^{\text{opt}}_{t+h|t} = \frac{\boldsymbol{\Sigma}^{-1}_{t+h|t}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}_{t+h|t}\mathbf{1}}, \tag{2}$$

where $\boldsymbol{\Sigma}_{t+h|t}$ is the $N \times N$ covariance matrix of the lead h forecast errors and $\mathbf{1}$ is the N-dimensional unit vector. Unfortunately, in practice, the elements of the covariance matrix $\boldsymbol{\Sigma}_{t+h|t}$ are usually unknown and required to be properly estimated.

It follows that if $\boldsymbol{w}_{t+h|t}$ is determined by Equation (2), one can identify a combined forecast $\tilde{y}_{t+h|t}$ with no greater error variance than the minimum error variance of all individual forecasts. The fact was further demonstrated in detail in Timmermann (2006) to illustrate the diversification gains offered by forecast combinations by simply considering the combination of two forecasts. Under mean squared error (MSE) loss, Timmermann (2006) characterized the general solution of the optimal linear combination weights given the joint Gaussian distribution of the outcome y_{t+h} and forecasts $\hat{\mathbf{y}}_{t+h|t}$.

The loss assumed in Bates and Granger (1969) and Newbold and Granger (1974) is quadratic and symmetric in the forecast error from the linear combination. Elliott and Timmermann (2004) examined forecast combinations under more general loss functions that account for asymmetries, and forecast error distributions with skew. They demonstrated that the optimal combination weights in a combination strongly depended on the degree of asymmetry in the loss function and skews in the underlying forecast error distribution. Subsequently, Patton and Timmermann (2007) demonstrated that the properties of optimal forecasts established under MSE loss were not generally robust under more general assumptions about the loss function. The properties of optimal forecasts were also generalized to consider asymmetric loss and nonlinear DGP.

Regression approach

The seminal work by Granger and Ramanathan (1984) provided an important impetus for approximating the optimal weights under a linear regression framework. They recommended the strategy that the combination weights can be estimated by Ordinary Least Squares (OLS) in regression models having the vector of past observations as the response variable and the matrix of past forecasts as the explanatory variables. Three alternative approaches involving various possible restrictions are considered

$$y_{t+h} = \boldsymbol{w}_h' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}, \quad s.t. \quad \boldsymbol{w}_h' \mathbf{1} = 1, \tag{3}$$

$$y_{t+h} = \boldsymbol{w}' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}, \tag{4}$$

$$y_{t+h} = \omega_{0h} + \boldsymbol{w}' \hat{\mathbf{y}}_{t+h|t} + \varepsilon_{t+h}. \tag{5}$$

The constrained OLS estimation of the regression (3) in which the constant is omitted and the weights are constrained to sum to one yields results identical to the optimal weights proposed by Bates and Granger (1969). Furthermore, Granger and Ramanathan (1984) suggested the unrestricted OLS regression (5) which allowed for a constant term and did not impose the weights sum to one was superior to the popular optimal method regardless of whether the constituent forecasts were biased. However, De Menezes et al. (2000) put forward some consideration required when using the unrestricted regression, including the stationarity of the series being forecast, the possible presence of serial correlation in forecast errors (see also Diebold, 1988; Edward Coulson and Robins, 1993), and the issue of multicollinearity.

More generalizations of the combination regressions have been considered in the literature. Diebold (1988) exploited the serial correlation in least squares framework by characterizing the combined forecast errors as the AutoRegressive Moving Average (ARMA) processes, leading to improved combined forecasts. Gunter (1992) and Aksu and Gunter (1992) provided an empirical analysis to compare the performance of various combination strategies, including the simple average, the unrestricted OLS regression, the restricted OLS regression where the weights were restricted to sum to unity, and the nonnegativity restricted OLS regression where the weights were constrained to be nonnegative. The results revealed that constraining weights to be nonnegative was at least as robust and accurate as the simple average and yielded superiority over other combinations based on regression framework. Conflitti et al. (2015) addressed the problem of determining the optimal weights by imposing two restrictions that the weights should be nonnegative and sum to one, which turned out to be a special case of a lasso regression. Edward Coulson and Robins (1993) found that allowing a lagged dependent variable in forecast combination regressions could achieve improved performance. Instead of using the quadratic loss function, Nowotarski et al. (2014) applied the absolute loss function in the unrestricted regression to yield the least absolute deviation regression which was more robust to outliers than OLS combinations.

The forecast combinations using changing weights are developed in the relevant literature to consider various types of structual changes in the constituent forecasts. For instance, Diebold and Pauly (1987) explored the possibilities for time-varying parameters in regression-based combination approaches. Both deterministic and stochastic time-varying parameters are considered in the linear regression framework. Specifically, the combination weights are described as deterministic nonlinear (polynomial) functions of time or allowed to involve random variation. Deutsch et al. (1994) allowed the combination weights to evolve immediately or smoothly using switching regression models and smooth transition regression models.

Researchers have worked on dealing with a large number of forecasts in the regression framework to take advantages of many different models. Chan et al. (1999) examined a wide range of combination methods in a Monte Carlo experiment and a real-word dataset. Their results investigated the poor performance of OLS combinations when the number of forecasts to be combined was large and suggested alternative weight estimation methods, such as ridge regression and principal components forecast combination. Stock and Watson (2004) offered the details of principal component forecast combination, which entailed forming a regression having the actual value as

the response variable and the first few principal components reduced from several forecasts as the explanatory variables. This method reduces the number of weights that must be estimated in a regression framework, and frequently serves as a way to solve the multicollinearity problem which is likely to lead to unstable behavior in the estimated weights. The superiority of the principal components regression that involved dimension reduction techniques over OLS combinations was also supproted in Rapach and Strauss (2008) and Poncela et al. (2011).

Performance-based weights

Estimation errors in the optimal weights and a diverse set of regression-based weights tend to be particularly large due to difficulties in properly estimating the entire covariance matrix $\mathbf{\Sigma}_{t+h|t}$, especially in situations with the large number of forecasts at hand. Instead, Bates and Granger (1969) suggested weighting the constituent forecasts in inverse proportion to their historical performance, ignoring correlations across forecast errors. In follow-up studies, Newbold and Granger (1974) and Winkler and Makridakis (1983) generalized the issue in the sense of considering more time series, more forecasting models, and multiple forecast horizons. Their extensive results demonstrated that combinations which took account of correlations performed poorly, and consequently reconfirmed Bates and Granger (1969) argument that correlations can be poorly estimated in practice and should be ignored in calculating combination weights.

Let $\mathbf{e}_{t+h|t} = \mathbf{1}y_{t+h} - \hat{\mathbf{y}}_{t+h|t}$ be the N-vector of h-period forecast errors from the individual models, the five procedures suggested in Bates and Granger (1969) for estimating the combination

weights when $\boldsymbol{\Sigma}_{t+h|t}$ is unknown, extended to the general case are as follows:

$$w^{\text{bg1}}_{t+h|t,i} = \frac{\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,j}\right)^{-1}}. \tag{6}$$

$$\boldsymbol{w}^{\text{bg2}}_{t+h|t} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}, \quad \text{where} \quad (\hat{\boldsymbol{\Sigma}}_{t+h|t})_{i,j} = \nu^{-1}\sum_{\tau=t-\nu+1}^{t} e_{\tau|\tau-h,i}e_{\tau|\tau-h,j}. \tag{7}$$

$$w^{\text{bg3}}_{t+h|t,i} = \alpha\hat{w}_{t+h-1|t-1,i} + (1-\alpha)\frac{\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=t-\nu+1}^{t} e^2_{\tau|\tau-h,j}\right)^{-1}}, \quad 0<\alpha<1. \tag{8}$$

$$w^{\text{bg4}}_{t+h|t,i} = \frac{\left(\sum_{\tau=1}^{t} \gamma^{\tau} e^2_{\tau|\tau-h,i}\right)^{-1}}{\sum_{j=1}^{N}\left(\sum_{\tau=1}^{t} \gamma^{\tau} e^2_{\tau|\tau-h,j}\right)^{-1}}, \quad \gamma \geq 1. \tag{9}$$

$$\boldsymbol{w}^{\text{bg5}}_{t+h|t} = \frac{\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}_{t+h|t}\mathbf{1}}, \quad \text{where} \quad (\hat{\boldsymbol{\Sigma}}_{t+h|t})_{i,j} = \frac{\sum_{\tau=1}^{t}\gamma^{\tau}e_{\tau|\tau-h,i}e_{\tau|\tau-h,j}}{\sum_{\tau=1}^{t}\gamma^{\tau}} \quad \text{and} \quad \gamma \geq 1. \tag{10}$$

These weighting schemes differ in the factors, as well as the choice of the parameters, ν, α, and γ. Correlations across forecast errors are either ignored by treating the covariance matrix $\boldsymbol{\Sigma}_{t+h|t}$ as a diagonal matrix or estimated using sample data points which, however, may lead to quite unstable estimates of $\boldsymbol{\Sigma}_{t+h|t}$ given highly correlated forecast errors. Some estimation schemes suggest computing or updating the relative performance of different models over rolling windows of the most recent ν observations, while others base the weights on exponential discounting with higher values of γ giving larger weights to recent observations. In consequence, these weighting schemes are well adapted to allow the non-stationary relationship between the individual forecasting procedures over time (Newbold and Granger, 1974), which, however, tends to increase the variance of the parameter estimates and works quite poorly provided that the DGP is truly covariance stationary (Timmermann, 2006).

A broader set of combination weights based on the relative performance of individual forecasting techniques is developed and examined in a series of studies. Stock and Watson (1998) generalized the rolling window scheme (6) in the sense that the weights on the individual forecasts were inversely proportional to the kth power of their MSE. The weights with $k = 0$ correspond to assigning equal weights to all forecasts, while more weights are placed on the best performing models by considering $k \geq 1$. Other forms of forecast error measures, such as the Root Mean Squared Error (RMSE) and the symmetric Mean Absolute Percentage Error (sMAPE), are also considered to

develop the performance-based combination weights (e.g., Nowotarski et al., 2014; Pawlikowski and Chorowska, 2020). Besides, a weighting scheme with the weights depending inversely on the exponentially discounted errors is proposed by Stock and Watson (2004) as an upgraded version of the scheme (9), and encompassed in the sequent studies (e.g., Clark and McCracken, 2010; Genre et al., 2013) to achieve gains from combining forecasts. The pseudo out-of-sample performance used in these weighting schemes is commonly computed based on rolling or recursive (expanding) windows (e.g., Stock and Watson, 1998; Clark and McCracken, 2010; Genre et al., 2013). It is natural to adopt rolling windows in estimating the weights to deal with the structural change. But the window length should not be too short without the estimates of the weights becoming too noisy (Baumeister and Kilian, 2015).

Compared to constructing the weights directly using historical forecast errors, a new form of combination that is more robust and less sensitive to outliers is introduced based on the 'ranking' of models. Again this combination ignores correlations across forecast errors. The simplest and most commonly used method in the class is to use the median forecast as the output. Aiolfi and Timmermann (2006) constructed the weights proportional to the inverse of performance ranks (sorted according to increasing order of forecast errors), which were later used by Andrawis et al. (2011) for tourism demand forecasting. Another weighting scheme that attaches a weight proportional to $\exp(\beta(N+1-i))$ to the ith ordered constituent model is adopted in Yao and Islam (2008) and Donate et al. (2013) to combine Artificial Neural Networks (ANNs), where β is a scaling factor. However, as mentioned by Andrawis et al. (2011), this class of combination method still comes with the drawback of the discrete nature because it limits the weight to only a few possible levels.

Combinations based on information criteria

Information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974), the corrected Akaike Information Criterion (AICc, Sugiura, 1978), and the Bayesian Information Criterion (BIC, Schwarz, 1978), are often advised to deal with model selection in forecasting. However, choosing a single model out of the candidate model pool may be misleading because of the loss of information gleaned from alternative models. An alternative way proposed by Burnham and Anderson (2002) is to combine different models based on information criteria to mitigate the risk of selecting a single model.

One such common approach is using Akaike weights. Specifically, in light of the fact that AIC estimates the Kullback-Leibler distance (Kullback and Leibler, 1951) between a model and the true DGP, differences in the AIC can be considered to weight different models, providing a measure of the evidence for each model relative to other constituent models. Given N individual models, the Akaike weight of model i can be derived by the following steps:

$$\Delta \text{AIC}_i = \text{AIC}_i - \min_{k \in \{1,2,\cdots,N\}} \text{AIC}(k), \tag{11}$$

$$w_i^{\text{aic}} = \frac{\exp(-0.5\Delta \text{AIC}_i)}{\sum_{k=1}^{N} \exp\left(-0.5\Delta \text{AIC}_k\right)}. \tag{12}$$

Akaike weights calculated in this manner can be interpreted as the probability that a given model performs best at approximating the unknown DGP, given the model set and data (Kolassa, 2011). Similar weights from AICc and BIC can be derived analogously to Equation (11)-(12).

The outstanding performance of weighted combinations based on information criteria has been confirmed in several research. For instance, Kolassa (2011) used weights derived from AIC, AICc and BIC to combine exponential smoothing forecasts, and resulted in superior accuracy over selection using these information criteria. The similar strategy was adopted by Petropoulos, Hyndman and Bergmeir (2018) to separately explore the benefits of bagging for time series forecasting. Furthermore, an empirical study by Petropoulos, Kourentzes, Nikolopoulos and Siemsen (2018) showed that a weighted combination based on AIC improved the performance of the statistical benchmark they used.

Bayesian approach

Some effort has been directed toward the use of Bayesian approaches to updating forecast combination weights in face of new information from various sources. Recall that obtaining reliable estimates of the covariance matrix $\mathbf{\Sigma}$ (the time and horizon subscripts are dropped for simplicity) of forecast errors with correlation being ignored or not, is a major challenge in the general case. With this in mind, Bunn (1975) suggested the idea of a Bayesian combination on the basis of the probability of respective forecasting model performing the best on any given occasion. Considering the beta and the Dirichlet distributions arising as the conjugate priors for the binomial and multinomial processes respectively, the suggested non-parametric method performs well when there is relatively little past data by the way of attaching prior subjective probabilities to individual

forecasts (Bunn, 1985; De Menezes et al., 2000). Öller (1978) presented another way to involve subjective probability in a Bayesian method based on the self-scoring weights proportional to the evaluation of the expert's forecasting ability.

A different theme of research has also advocated the incorporation of prior information into the estimation of combination weights, but with the weights being shrunk toward some prior mean (Newbold and Harvey, 2002). Assuming that forecast errors are jointly normally distributed, Clemen and Winkler (1986) developed a Bayesian approach with the conjugate prior for $\mathbf{\Sigma}$, represented by an inverted Wishart distribution with covariance matrix $\mathbf{\Sigma}_0$ and scalar degrees of freedom ν_0. Again we drop time and horizon subscripts for simplicity. If the last n observations are used to estimate $\mathbf{\Sigma}$, the combination weights derived from the posterior distribution for $\mathbf{\Sigma}$ is

$$\boldsymbol{w}^{\text{cw}} = \frac{\mathbf{\Sigma}^{*-1}\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}^{*-1}\mathbf{1}}, \tag{13}$$

where the covariance matrix $\mathbf{\Sigma}^* = \left[\left(\nu_0\mathbf{\Sigma}_0^{-1} + n\hat{\mathbf{\Sigma}}^{-1}\right)/(\nu_0 + n)\right]^{-1}$ and $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix. The prior estimate $\mathbf{\Sigma}_0$ can be specified to allow correlations across forecast errors. The subsequent work by Diebold and Pauly (1990) considered a normal regression-based combination $\mathbf{y} = \hat{\mathbf{y}}\boldsymbol{w} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is distributed as $N\left(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}\right)$, with the standard normal-gamma conjugate prior. The approach results in the estimated combination weights which can be considered as a matrix weighted average of those for the two polar cases, least squares and prior weights. Due to the fact that Bayesian approaches have been mostly employed to construct combinations of probability forecasts, we will elaborate on other methods of determining combination weights in a Bayesian context in the following Section 3.

2.2.2 Nonlinear combinations

2.2.3 Combining by learning

- Stacking: introduce stacking as a combination/aggregation method rather than a generalization of many ensemble methods in ML.
- Regression-based combinations: a simple stacking.
- Meta-learning that only takes individual forecasts as input.

- Stacking extension: use other information in the meta-learner, such as feature and diversity (FFORMA).

2.2.4 Which forecasts should be combined?

forecast pooling / optimal forecast groups / model selection (Zhou et al., 2002; Kourentzes et al., 2019)

2.3 Forecast Combination Puzzle

(why don't weights work?)

(Kang, 1986; De Menezes et al., 2000; Genre et al., 2013; Post et al., 2019; Chan and Pauwels, 2018; Lichtendahl and Winkler, 2020; Kourentzes et al., 2019)

Diebold and Pauly (1990) uses Bayesian shrinkage techniques to allow the incorporation of prior information into the estimation of combining weights; least squares and prior weights then emerge as polar cases for the posterior mean.

3 Probabilistic forecast combinations

- Combining quantiles and prediction intervals.
- Combining distributions as in fable.

(Genest and McConway, 1990; Nowotarski and Weron, 2015; Conflitti et al., 2015; Billio et al., 2013)

Bayesian analysis (Winkler, 1968; Clemen and Winkler, 1999)

BMA (Li et al., 2011)

4 Probabilistic ensembles

- Meteorological ensembles.

- True ensembles in other areas (i.e., not papers that use the word "ensemble" but papers that use mixtures when forecasting).
- When is an ensemble equivalent to combination?
- When do point forecasts from an ensemble equal point forecasts from a combination?

5 Boosting in forecasting

6 Bagging in forecasting

7 Stacking in forecasting

References

Agnew, C. E. (1985), 'Bayesian consensus forecasts of macroeconomic variables', *Journal of Forecasting* **4**(4), 363–376.

Aiolfi, M. and Timmermann, A. (2006), 'Persistence in forecasting performance and conditional combination strategies', *Journal of Econometrics* **135**(1), 31–53.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aksu, C. and Gunter, S. I. (1992), 'An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts', *International Journal of Forecasting* **8**(1), 27–43.

Andrawis, R. R., Atiya, A. F. and El-Shishiny, H. (2011), 'Combination of long term and short term forecasts, with application to tourism demand forecasting', *International Journal of Forecasting* **27**(3), 870–886.

Atiya, A. F. (2020), 'Why does forecast combination work so well?', *International Journal of Forecasting* **36**(1), 197–200.

Barnard, G. A. (1963), 'New methods of quality control', *Journal of the Royal Statistical Society. Series A* **126**(2), 255.

Bates, J. M. and Granger, C. W. J. (1969), 'The combination of forecasts', *The Journal of the Operational Research Society* **20**(4), 451–468.

Baumeister, C. and Kilian, L. (2015), 'Forecasting the real price of oil in a changing world: A forecast combination approach', *Journal of Business & Economic Statistics* **33**(3), 338–351.

Billah, B., Hyndman, R. J. and Koehler, A. B. (2005), 'Empirical information criteria for time series forecasting model selection', *Journal of Statistical Computation and Simulation* **75**(10), 831–840.

Billio, M., Casarin, R., Ravazzolo, F. and van Dijk, H. K. (2013), 'Time-varying combinations of predictive densities using nonlinear filtering', *Journal of Econometrics* **177**(2), 213–232.

Bunn, D. W. (1975), 'A bayesian approach to the linear combination of forecasts', *Journal of the Operational Research Society* .

Bunn, D. W. (1985), 'Statistical efficiency in the linear combination of forecasts', *International Journal of Forecasting* **1**(2), 151–163.

Burnham, K. P. and Anderson, D. R. (2002), 'Model selection and multi-model inference: A practical information-theoretic approach (2nd ed.)', *Berlin, New York: Springer* .

Chan, F. and Pauwels, L. L. (2018), 'Some theoretical results on forecast combinations', *International Journal of Forecasting* **34**(1), 64–74.

Chan, Y. L., Stock, J. H. and Watson, M. W. (1999), 'A dynamic factor model framework for forecast combination', *Spanish Economic Review* **1**(2), 91–121.

Cheng, X. and Hansen, B. E. (2015), 'Forecasting with factor-augmented regression: A frequentist model averaging approach', *Journal of Econometrics* **186**(2), 280–293.

Clark, T. E. and McCracken, M. W. (2010), 'Averaging forecasts from VARs with uncertain instabilities', *Journal of Applied Econometrics* .

Clemen, R. T. (1989), 'Combining forecasts: A review and annotated bibliography', *International Journal of Forecasting* **5**(4), 559–583.

Clemen, R. T. and Winkler, R. L. (1986), 'Combining economic forecasts', *Journal of Business & Economic Statistics* **4**(1), 39–46.

Clemen, R. T. and Winkler, R. L. (1999), 'Combining probability distributions from experts in risk analysis', *Risk Analysis* **19**(2), 187–203.

Conflitti, C., De Mol, C. and Giannone, D. (2015), 'Optimal combination of survey forecasts', *International Journal of Forecasting* **31**(4), 1096–1103.

Crane, D. B. and Crotty, J. R. (1967), 'A two-stage forecasting model: Exponential smoothing and multiple regression', *Management Science* **13**(8), B–501.

De Menezes, L. M., Bunn, D. W. and Taylor, J. W. (2000), 'Review of guidelines for the use of combined forecasts', *European Journal of Operational Research* .

Deutsch, M., Granger, C. W. J. and Teräsvirta, T. (1994), 'The combination of forecasts using changing weights', *International Journal of Forecasting* **10**(1), 47–57.

Diebold, F. X. (1988), 'Serial correlation and the combination of forecasts', *Journal of Business & Economic Statistics* **6**(1), 105–111.

Diebold, F. X. and Pauly, P. (1987), 'Structural change and the combination of forecasts', *Journal of Forecasting* **6**(1), 21–40.

Diebold, F. X. and Pauly, P. (1990), 'The use of prior information in forecast combination', *International Journal of Forecasting* **6**(4), 503–508.

Donate, J. P., Cortez, P., Sanchez, G. G. and De Miguel, A. S. (2013), 'Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble', *Neurocomputing* .

Edward Coulson, N. and Robins, R. P. (1993), 'Forecast combination in a dynamic setting', *Journal of Forecasting* **12**(1), 63–67.

Elliott, G. and Timmermann, A. (2004), 'Optimal forecast combinations under general loss functions and forecast error distributions', *Journal of Econometrics* **122**(1), 47–79.

Fildes, R. and Petropoulos, F. (2015), 'Simple versus complex selection rules for forecasting many time series', *Journal of Business Research* **68**(8), 1692–1701.

Genest, C. and McConway, K. J. (1990), 'Allocating the weights in the linear opinion pool', *Journal of Forecasting* **9**(1), 53–73.

Genre, V., Kenny, G., Meyler, A. and Timmermann, A. (2013), 'Combining expert forecasts: Can anything beat the simple average?', *International Journal of Forecasting* **29**(1), 108–121.

Granger, C. W. J. (1989), 'Invited review combining forecasts—twenty years later', *Journal of Forecasting* **8**(3), 167–173.

Granger, C. W. J. and Ramanathan, R. (1984), 'Improved methods of combining forecasts', *Journal of Forecasting* **3**(2), 197–204.

Grushka-Cockayne, Y., Jose, V. R. R. and Lichtendahl, K. C. (2017), 'Ensembles of overfit and overconfident forecasts', *Management Science* **63**(4), 1110–1130.

Gunter, S. I. (1992), 'Nonnegativity restricted least squares combinations', *International Journal of Forecasting* **8**(1), 45–59.

Inoue, A., Jin, L. and Rossi, B. (2017), 'Rolling window selection for out-of-sample forecasting with time-varying parameters', *Journal of Econometrics* **196**(1), 55–67.

Jose, V. R. R., Grushka-Cockayne, Y. and Lichtendahl, K. C. (2014), 'Trimmed opinion pools and the crowd's calibration problem', *Management Science* **60**(2), 463–475.

Jose, V. R. R. and Winkler, R. L. (2008), 'Simple robust averages of forecasts: Some empirical results', *International Journal of Forecasting* **24**(1), 163–169.

Kang, H. (1986), 'Unstable weights in the combination of forecasts', *Management Science* **32**(6), 683–695.

Kang, Y., Hyndman, R. J. and Li, F. (2020), 'GRATIS: GeneRAting TIme Series with diverse and controllable characteristics', *Statistical Analysis and Data Mining* **13**(4), 354–376.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'IJCAI', Vol. 14, Montreal, Canada, pp. 1137–1145.

Kolassa, S. (2011), 'Combining exponential smoothing forecasts using akaike weights', *International Journal of Forecasting* **27**(2), 238–251.

Kourentzes, N., Barrow, D. K. and Crone, S. F. (2014), 'Neural network ensemble operators for time series forecasting', *Expert Systems with Applications* **41**(9), 4235–4244.

Kourentzes, N., Barrow, D. and Petropoulos, F. (2019), 'Another look at forecast selection and combination: Evidence from forecast pooling', *International Journal of Production Economics* **209**(February 2018), 226–235.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* .

Leutbecher, M. and Palmer, T. N. (2008), 'Ensemble forecasting', *Journal of Computational Physics* **227**(7), 3515–3539.

Lewis, J. M. (2005), 'Roots of ensemble forecasting', *Monthly Weather Review* **133**(7), 1865–1885.

Li, G., Shi, J. and Zhou, J. (2011), 'Bayesian adaptive combination of short-term wind speed forecasts from neural network models', *Renewable Energy* **36**(1), 352–359.

Lichtendahl, K. C. and Winkler, R. L. (2020), 'Why do some combinations perform better than others?', *International Journal of Forecasting* **36**(1), 142–149.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982), 'The accuracy of extrapolation (time series) methods: Results of a forecasting competition', *Journal of Forecasting* **1**(2), 111–153.

Makridakis, S. and Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* **16**(4), 451–476.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020), 'The M4 competition: 100,000 time series and 61 forecasting methods', *International Journal of Forecasting* **36**(1), 54–74.

Makridakis, S. and Winkler, R. L. (1983), 'Averages of forecasts: Some empirical results'.

McNees, S. K. (1992), 'The uses and abuses of 'consensus' forecasts', *Journal of Forecasting* **11**(8), 703–710.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J. and Talagala, T. S. (2020), 'FFORMA: Feature-based forecast model averaging', *International Journal of Forecasting* **36**(1), 86–92.

Newbold, P. and Granger, C. W. J. (1974), 'Experience with forecasting univariate time series and the combination of forecasts', *Journal of the Royal Statistical Society. Series A* **137**(2), 131.

Newbold, P. and Harvey, D. I. (2002), 'Forecast combination and encompassing', *A companion to economic forecasting* .

Nowotarski, J., Raviv, E., Trück, S. and Weron, R. (2014), 'An empirical comparison of alternative schemes for combining electricity spot price forecasts', *Energy Economics* **46**, 395–412.

Nowotarski, J. and Weron, R. (2015), 'Computing electricity spot price prediction intervals using quantile regression and forecast averaging', *Computational Statistics* **30**(3), 791–803.

Öller, L.-E. (1978), 'A method for pooling forecasts', *Journal of the Operational Research Society* **29**(1), 55–63.

Palm, F. C. and Zellner, A. (1992), 'To combine or not to combine? issues of combining forecasts', *Journal of Forecasting* **11**(8), 687–701.

Patton, A. J. and Timmermann, A. (2007), 'Properties of optimal forecasts under asymmetric loss and nonlinearity', *Journal of Econometrics* **140**(2), 884–918.

Pawlikowski, M. and Chorowska, A. (2020), 'Weighted ensemble of statistical models', *International Journal of Forecasting* **36**(1), 93–97.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Bergmeir, C., Bessa, R. J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., Fiszeder, P., Franses, P. H., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A. and Ziel, F. (2020), 'Forecasting: theory and practice'.

Petropoulos, F., Hyndman, R. J. and Bergmeir, C. (2018), 'Exploring the sources of uncertainty: Why does bagging for time series forecasting work?', *European Journal of Operational Research* **268**(2), 545–554.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K. and Siemsen, E. (2018), 'Judgmental selection of forecasting models', *Journal of Operations Management* **60**, 34–46.

Petropoulos, F. and Svetunkov, I. (2020), 'A simple combination of univariate models', *International Journal of Forecasting* **36**(1), 110–115.

Poler, R. and Mula, J. (2011), 'Forecasting model selection through out-of-sample rolling horizon weighted errors', *Expert Systems with Applications* **38**(12), 14778–14785.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R. and Senra, E. (2011), 'Forecast combination through dimension reduction techniques', *International Journal of Forecasting* **27**(2), 224–237.

Post, T., Karabatı, S. and Arvanitis, S. (2019), 'Robust optimization of forecast combinations', *International Journal of Forecasting* **35**(3), 910–926.

Qi, M. and Zhang, G. P. (2001), 'An investigation of model selection criteria for neural network time series forecasting', *European Journal of Operational Research* **132**(3), 666–680.

Rapach, D. E. and Strauss, J. K. (2008), 'Forecasting US employment growth using forecast combining methods', *Journal of Forecasting* **27**(1), 75–93.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**(2), 461–464.

Shaub, D. (2019), 'Fast and accurate yearly time series forecasting with forecast combinations', *International Journal of Forecasting* .

Stock, J. H. and Watson, M. W. (1998), A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, Technical Report w6607, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2003), How did leading indicator forecasts perform during the 2001 recession?

Stock, J. H. and Watson, M. W. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**(6), 405–430.

Sugiura, N. (1978), 'Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's', *Communications in Statistics-Theory and Methods* .

Talagala, T. S., Hyndman, R. J. and Athanasopoulos, G. (2018), 'Meta-learning how to forecast time series', *Monash Econometrics and Business Statistics Working Papers* **6**, 18.

Timmermann, A. (2006), Chapter 4 forecast combinations, *in* G. Elliott, C. W. J. Granger and A. Timmermann, eds, 'Handbook of Economic Forecasting', Vol. 1, Elsevier, pp. 135–196.

Wang, X., Kang, Y., Petropoulos, F. and Li, F. (2021), 'The uncertainty estimation of feature-based forecast combinations', *Journal of the Operational Research Society* .

Winkler, R. L. (1968), 'The consensus of subjective probability distributions', *Management Science* **15**(2), B–61–B–75.

Winkler, R. L. and Makridakis, S. (1983), 'The combination of forecasts', *Journal of the Royal Statistical Society. Series A* **146**(2), 150.

Yang, Y. (2005), 'Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation', *Biometrika* **92**(4), 937–950.

Yao, X. and Islam, M. M. (2008), 'Evolving artificial neural network ensembles', *IEEE Computational Intelligence Magazine* **3**(1), 31–42.

Zhou, Z.-H., Wu, J. and Tang, W. (2002), 'Ensembling neural networks: Many could be better than all', *Artificial Intelligence* **137**(1), 239–263.