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, 1998) and the Bayesian Information Criterion (BIC, Schwarz et al., 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 is 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.

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 can be substantially improved by combining multiple forecasts. Timmermann (2006) summarized the benefits of forecast combinations into the fact that individual forecasts are 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 are beneficial due to diversification gains. More recently, Atiya (2020) illustrated graphically why forecast combinations are superior.

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 (see Clemen, 1989). 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—that is, simple combination schemes do not require estimating parameters such as combination weights based on forecast errors, thus avoiding parameter estimation error that often exists in the weighted combination.

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 is the most widely used simple combination rule (see Bunn, 1985; Clemen and Winkler, 1986; Stock and Watson, 2003). Makridakis et al. (1982) reported the results of M forecasting competition and found that a simple average outperformed all individual methods. Clemen (1989) provided a review and annotated bibliography of the early work on the combination of forecasts, and then addressed the issue that the arithmetic means often dominate more refined forecast combinations. Makridakis and Winkler (1983) shown that combining forecasts using simple average reduces the variability of accuracy and hence the risk associated with selecting the best forecast. Palm and Zellner (1992) concisely summarized the advantages of adopting a simple average into three points: (i) combination weights are equal and do not have to be estimated, (ii) a simple average significantly reduces 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 is taken into account. Furthermore, Timmermann (2006) pointed out that the good average performance of the simple average depends 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 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). McNees (1992) found no significant difference between the mean and the trimmed mean, while the results of Stock and Watson (2004) support the mean. 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 are 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 deal 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 have 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) show that simple combinations can 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). Therefore, simple combination rules 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 Combination Weighting Schemes

2.1 Regression-based combination

2.2 Relative performance weights

2.3 Criterion-based combination

2.4 Combining by learning

2.5 Time-varying combination

2.6 Nonlinear combination

3 Forecast Combination Puzzle

4 Probabilistic Forecast Combinations

References

Akaike, H. (1998), Information theory and an extension of the maximum likelihood principle, *in* 'Selected papers of hirotugu akaike', Springer, pp. 199–213.

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.

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.

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

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.

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.

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

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

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.

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

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, 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.

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.

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.

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.

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

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. 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.

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.

Schwarz, G. et al. (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. (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.

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* .

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.