Journal of Business Research 69 (2016) 3951–3962

Contents lists available at ScienceDirect

Journal of Business Research

When to choose the simple average in forecast combination

Sebastian M. Blanc *, Thomas Setzer

Karlsruhe Institute of Technology, Fritz-Erler-Straße 23, 76131 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 17 May 2016
Accepted 21 May 2016
Available online xxxx

Keywords:
Forecast combination
Simple average
Optimal weights
Structural changes

ABSTRACT

Numerous forecast combination techniques have been proposed. However, these do not systematically outperform a simple average (SA) of forecasts in empirical studies. Although it is known that this is due to instability of learned weights, managers still have little guidance on how to solve this "forecast combination puzzle", i.e., which combination method to choose in specific settings. We introduce a model determining the yet unknown asymptotic out-of-sample error variance of the two basic combination techniques: SA, where no weightings are learned, and so-called optimal weights that minimize the in-sample error variance. Using the model, we derive multi-criteria boundaries (considering training sample size and changes of the parameters which are estimated for optimal weights) to decide when to choose SA. We present an empirical evaluation which illustrates how the decision rules can be applied in practice. We find that using the decision rules is superior to all other considered combination strategies.

1. Introduction

The combination of forecasts has been subject to research in economics since the pioneering work of Reid (1968) and Bates and Granger (1969). Numerous studies show that the combination of forecasts often results in increased accuracy in comparison to any of the forecasts alone (Makridakis et al., 1982; Clemen, 1989; Makridakis & Hibon, 2000; Fildes & Petropoulos, 2015). Various techniques aiming at deriving a weighting of individual forecasts which minimizes errors out-of-sample have been proposed.

Bates and Granger (1969) introduced the so-called optimal weights (OW). The weights are determined in a least squares estimation using available past forecast error data. They are referred to as optimal as they minimize the in-sample error variance; by design, OW outperforms any other linear weighting approach in-sample. However, the out-of-sample performance is not necessarily superior since the estimated weights are strongly fitted to the training data and are consequently subject to sampling-based variance.

As a consequence, alternative weight estimation approaches have been proposed. Clemen (1989); Diebold and Lopez (1996), and Timmermann (2006) provided thorough literature reviews of the various approaches to forecast combination. Approaches include variants of optimal weights constrained to the interval [0,1], shrinkage towards the average, Bayesian outperformance probabilities, and several more approaches. Each of the alternative approaches outperformed OW as well as other approaches out-of-sample in some evaluations, but are outperformed in others. As no model exists to decide which of the approaches to choose and empirical results are ambiguous, there is no clear consensus on which forecast combination method can be expected to perform best in a particular situation.

A surprising observation of the reviews was, however, that amongst the approaches under study, the simple average (SA) was not systematically outperformed by any other approach in out-of-sample evaluations. Stock and Watson (2004) coined the term "forecast combination puzzle" for this phenomenon. Besides model-based forecasting, SA is also competitive when combining expert predictions. For instance Genre, Kenny, Meyler, and Timmermann (2013) found that for forecasts of unemployment rate and GDP growth, only few combination methods outperform SA, while their results caution against any assumption that the identified improvements would persist in the future.

The forecast combination puzzle is in line with the more general phenomenon that simpler forecasting procedures usually outperform more complex techniques. Green and Armstrong (2015) reviewed 97 studies comparing simple and complex methods, concluding that "none of the papers provide a balance of evidence that complexity improves the accuracy of forecasts out-of-sample". Simplicity in forecasting procedures corresponds to using models where few different cues are used and/or few parameters have to be estimated. Likewise, in forecast combination, where weights of forecasts instead of cues are chosen, SA is the simplest model as – in contrast to more complex models such as OW – no parameters are estimated at all.

Brighton and Gigerenzer (2015) argued that the benefits of simplicity are often overlooked because of a "bias bias", where the importance of the bias component of the error is inflated. In contrast, the variance component, resulting from oversensitivity to different samples from the same population, is often ignored. Simpler approaches are typically

* Corresponding author.
E-mail addresses: sebastian.blanc@kit.edu (S.M. Blanc), thomas.setzer@kit.edu (T. Setzer).

http://dx.doi.org/10.1016/j.jbusres.2016.05.013
0148-2963/

more robust against different samples as the variance component is directly related to model complexity.

Simple averaging strategies have also been shown to be highly competitive in applications besides forecast combination. For instance, for venture capital decisions, Woike, Hoffrage, and Petty (2015) found that the decision quality when using equally weighted binary cues is comparable to more complex strategies, but even more robust. Graefe (2015) argued that estimating coefficients (weights) of predictors in multivariate models is only reasonable for large and reliable datasets and few predictors. For small and noisy datasets and a large number of predictors, the authors argued that including all relevant variables is more important than the weighting.

In forecast combination, the robustness of SA has been an important research topic and a considerable body of literature examines the forecast combination puzzle theoretically and empirically. As will be discussed in Section 2, results indicate that the robustness of SA stems from unstable weight estimates from small training samples or diverging forecast error characteristics between the training and the evaluation samples. In a broader sense, these findings support the "'Golden Rule of Forecasting", stating that forecasts are to be conservative (Armstrong, Green, & Graefe, 2015). That is because increasing asymmetry of weights results in higher sensitivity to the results of one individual forecast that is less counterbalanced by others.

Although these qualitative relations are known, managers still have little guidance on which method to choose in a particular setting. More specifically, we are not aware of any comprehensive quantitative decision guidance on when to choose OW or SA.

In this paper, we introduce a model for the expected out-of-sample error variance of a forecast combination, in particular when using SA and OW. Using the model, we derive multi-criteria decision boundaries determining whether OW or SA will lead to lower asymptotic error variance in a specific setting. Practitioners can furthermore use the thresholds to assess the robustness of a decision. We show that existing empirical guidelines can largely be explained by the model. Furthermore, in an empirical study with data from the M3 competition, we demonstrate that the recommendations and the thresholds can be used to implement successful combination decision strategies in practical settings.

2. Related work

A substantial amount of research has been conducted on the performance and robustness of SA in comparison to other forecast combination methods. A basic and intuitive finding is that the performance of SA depends on the ratio of the error variances of the forecasts as well as on their correlation. SA can be expected to perform well in case of similar error variances and low or medium error correlations (Bunn, 1985; Gupta & Wilton, 1987), since the weights which are optimal in the evaluation sample then approach equal weights. However, as shown by Dickinson (1973); Winkler and Clemen (1992), and Smith and Wallis (2009), SA can outperform other methods even for differing error variances or strongly correlated errors because of instable weight estimates. Elliott (2011) found that gains from using OW instead of SA are often too small to balance estimation errors. Claeskens, Magnus, Vasnev, and Wang (2016) showed that weight estimation can even introduce biases in combinations of unbiased forecasts.

Monte Carlo simulations by Kang (1986) and Gupta and Wilton (1987) confirmed that unstable weight estimates are key to the high competitiveness of SA. Evaluations on real-world data, for instance for U.S. money supply forecasts (Figlewski & Urich, 1983) or GNP forecasts (Kang, 1986; Clemen & Winkler, 1986) showed similar results.

Some guidelines to help decision-makers in selecting a combination method have been proposed. In the case of two forecasts, Schmittlein, Kim, and Morrison (1990) recommended SA for small sample sizes and for errors with similar variances and weak correlation. De Menezes, Bunn, and Taylor (2000) recommended SA only for approximately equal error variances and OW for large samples and low error correlation. In other cases, they suggested using outperformance probabilities (with small samples and unequal error variances), optimal weights constrained to the interval [0,1] (with medium or large samples and correlation over 0.5), or OW calculated with a correlation of zero instead of the estimated correlation, i.e., assuming uncorrelated errors (with medium sample sizes and correlations below 0.5). Thresholds for similarity/dissimilarity of error variances and sample size were, however, not quantified.

Both guidelines assume equal characteristics (error variances and covariances) of known training and unknown (future) observations. However, these characteristics might change over time because of structural changes in time series, which might influence the performance of OW and SA very differently. Miller, Clemen, and Winkler (1992) showed that SA can, in comparison to OW and other approaches, benefit from several types of structural breaks such as location shifts. Diebold and Pauly (1987) found that structural changes generally tend to impact complex approaches more than simpler ones as the estimated weights tend to increasingly differ from the ones that would minimize error in the evaluation sample.

In this paper, in contrast to existing guidelines, we propose an analytical model to determine whether SA will asymptotically outperform OW in a specific setting. We derive decision rules based on statistical considerations that do not only consider sample size and variance/covariance estimates, but also how much those values are allowed to divergence between training and evaluation sample for a decision to stay optimal. These thresholds are key to assessing the robustness of a decision but have received scant attention in the literature so far.

3. Forecast combination

Given two forecasts $\hat{y}_A$ and $\hat{y}_B$ for an event y, a combined forecast can be calculated by weighting both forecasts. The most common approach is a linear combination of the forecasts using weight w to derive a novel forecast $\hat{y}_C = w\hat{y}_A + (1-w)\hat{y}_B$. Assuming unbiased individual forecasts with errors $e_A = y - \hat{y}_A \sim \mathcal{N}(0, \sigma_A^2)$, $e_B = y - \hat{y}_B \sim \mathcal{N}(0, \sigma_B^2)$ and a correlation ρ between e_A and e_B, Bates and Granger (1969) proposed optimal weights (OW) minimizing the error variance of $\hat{y}_C$ in-sample. The original definition as well as an alternative one using the ratio of error standard deviations $\phi = \sigma_A/\sigma_B$ and the assumption $\sigma_A = 1$ (which, in combination, does not change the estimate) are presented in Eq. (1).

$$w = \frac{\sigma_B^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B} = \frac{1-\rho\phi}{1+\phi^2-2\rho\phi} \tag{1}$$

The in-sample error variance of a forecast combination with different weights is illustrated in Fig. 1. The individual error variances $\sigma_A^2 = 1$ and $\sigma_B^2 = 4$ are indicated by the dotted horizontal lines. The graph shows the error variance resulting from combining forecasts with OW, SA, and with static weights set to $1/\phi$ (2:1 in the example).

When using OW, the combined error variance never exceeds the lower of the two error variances. In contrast, the combined error variances with SA and a static 2:1 weighting are lowest for an error correlation of -1 and linearly increase with error correlation. At some level of error correlation, the combined error variance exceeds the one of the better forecast ($\sigma_A^2 = 1$, in our case). However, the combined error variance still never exceeds the higher error variance — in our case $\sigma_B^2 = 4$. In summary, the difference between error variance with fixed weights (SA or 2:1) and OW is small for strong negative correlations and strictly increases with error correlation.

While OW combination leads to lower in-sample error variance than any other weighting scheme (especially weightings that ignore error variances and error correlation in the training data), we reconsider that SA often outperforms OW out-of-sample, indicating that the estimated weights do not always fit unknown observations well.