Why Does Forecast Combination Work so Well?

Amir F. Atiya
Department of Computer Engineering
Cairo University
Giza, Egypt
amir@alumni.caltech.edu

January 23, 2019

Abstract

Forecast combinations were big winners at the M4 competition. In this note we reflect and analyze the reason for the success of forecast combination. We illustrate graphically how and in what cases do forecast combination produce good results. We also study the effect of forecast combination on the bias and the variance of the forecast.

1 Introduction

The M4 competition (Makridakis, Spiliotis, and Assimakopoulos [12]) is the latest in a series of M-competitions (Makridakis et al [10], [11]) that started several decades ago. These competitions are very beneficial in gaining in depth knowledge about the performance of the different forecasting models, and in drawing up best practices for forecasting. The M4 competition reinforced some of the knowledge gained from previous competitions, such as the general superiority of simple forecasting models. However, there are some new lessons learned from this competition. The top model is a hybrid approach, which utilizes both machine learning and statistical approaches.
Perhaps hybrid statistical/nonlinear models would be a promising direction to pursue. Another finding is that forecast combinations have been very successful. Of the 17 most accurate methods, 12 were combinations of forecasting models. Even though forecast combination is known to be a winning strategy to follow (Armstrong [2] and Clemen [4]), it is the first competition where it demonstrated such strong dominance. In this note we consider the topic of forecast combination, and contemplate the reasons for its success. We present some analysis that backs up the empirical research that has been performed on this topic.

Let \( u_i, i = 1, \ldots, N \) be the \( N \) forecasting models to be combined. The forecast combination can be written as

\[
\hat{u} = \sum_{i=1}^{N} w_i u_i
\]

where \( w_i \) is a combination weight. Usually, but not always, researchers use a convex combination, i.e. \( 0 \leq w_i \leq 1 \) and \( \sum_{i=1}^{N} w_i = 1 \). The motivation behind forecast combination is the fact that typical forecasting problems possess a small or rather finite history of points. So it is practically not possible to obtain the correct specification of the underlying data generation process. To hedge against the resulting inaccuracy of the derived forecasting model, it is beneficial to consider several forecasting models, and combine their forecasts. Irrespective of any estimation errors, there is another view of why forecast combination works well, and this is illustrated graphically next section.

\section{Graphical Illustration}

The following illustration explains why and in what situations are forecast combinations superior. Let the vector \( y = (y_1, \ldots, y_H)^T \) be the true time series values for the horizon to be forecasted. Consider that there are four candidate forecasting models, and assume that they produce forecast vectors \( u, v, q, \) and \( z \) for the considered horizon. Refer to Figure 1 for an illustration for a two-dimensional case (i.e. the \( y \) vector and the forecast vectors are 2-D). The forecast root mean square (RMS) error for some forecast \( u \) is given by:
\[ E = \sqrt{\frac{1}{H} \sum_{i=1}^{H} (u_i - y_i)^2} \]  \hspace{1cm} (2)

\[ = \frac{1}{\sqrt{H}} \text{distance}(u, y) \]  \hspace{1cm} (3)

Thus the distance in the graph is proportional to the RMS error. Forecasts \( u \) and \( v \) possess diversity, and therefore are located at different parts of the space (see Figure). Moreover, they have comparable distances from \( y \), and hence they have comparable RMS error. The line connecting them (modeled as \( wu + (1 - w)v \), where \( w \) is a weight between 0 and 1) corresponds to the location of the combination of these forecasts. One can observe that the distance between any point on the line and \( y \) is smaller than the distance between either \( u \) or \( v \) and \( y \). This indicates that the RMS error of the combined forecasts is better than that of either constituent forecast. This is especially pronounced at the midpoint of the line, which corresponds to equal weight combination.

Let us now look at the combination of the forecasts of points \( u \) and \( q \) (see Figure). These forecasts are very similar (or rather highly correlated), and comparable in error. From the distance analysis of the line between \( u \) and \( q \) and point \( y \) one can observe that the forecast combination does not add much value. But is also not harmful.

Consider now the combination of vectors \( u \) and \( z \). One can see that the distance between \( z \) and \( y \) is larger than that between \( u \) and \( y \), indicating that \( z \) is an inferior forecast. What even adds extra harm is that \( z \) is not diverse from \( u \), and is highly correlated. The line between \( u \) and \( z \) (corresponding to their forecast combination) possesses larger distance from \( y \), hence worse performance than that of \( u \) alone. The forecast \( z \) is therefore considered a drag on the performance of the forecast combination.

From these observations we can deduce the following facts, which are also verified in empirical research and in most time series forecasting competitions.

- Forecast combination should be a winning strategy if the constituent forecasts are either diverse or comparable in performance.

- One should exclude forecasts that are considerably worse than the best ones in the pool, unless they are very diverse from the rest. This also agrees with the recommendations discussed in the literature, see
Figure 1: An illustration of forecast combinations: The solid lines between the points (forecasts) $u, v, q, z$ represent the forecast combinations, and the dotted line to the true value $y$ is proportional to the RMS error.
Armstrong [2], and Timmermann [13]. Also Kourentzes, Barrow, and Petropoulos [9] make the point that one has to be selective for the pool of forecasts to be used for forecast combination.

- For example, one can include in the combination pool top forecasts that are within a 10% range in performance (for example SMAPE range from 15% to 16.5%). One may include forecasts that are 20% or 30% worse (e.g. an SMAPE of 18% to 20%), provided that they possess the needed diversity.

3 Diversity and Forecast Combination

Diversity in forecasts can be achieved by relying on variables of different sources. These could operate in different “pathways” in the way they affect the variable to be forecasted (see De Menezes, Bunn, and Taylor [5]). For example one could have a stock price forecast model based on purely the stock price time series. Another model could be based on fundamental company information (for example financial statement data, such as sales, earnings, debt, etc). The two forecasts would be considered diverse, because they are based on quantities with little direct relation. For another example consider GDP forecast based on economic indicators, and another GDP forecast based on the interest rate across the yield curve. Both of them are fairly diverse, and it would be a good idea to pool these forecasts.

Kourentzes et al [7], [8] and Andrawis, Atiya, and El-Shishiny [1] suggested a novel way to impart diversity by combining forecasts at different time scales. For example we could have a monthly forecast, time aggregate the data and produce a quarterly and/or a yearly forecast. Short term, medium term, and long term dynamics are influenced by different factors, and therefore bring in different information.

The only pitfall of forecast combination is when the data are rife with outliers, or has a very heavy tailed distribution. In such a case, one may want to limit the pool of forecasts to be combined to be just a small number. The reason is that a large pool will give more of a chance for an outlier forecast to creep into the pool, and ruin the combined forecast. An analysis into the “exotic” world of heavy tailed distributions, their effects on prediction models, and the numerous paradoxes they pose can be found in the work of Yousef and Kundu [14].
4 Forecast combination and the Bias-Variance Decomposition

The bias variance decomposition is a fundamental concept in all estimation problems, including forecasting, classification, and parameter estimation problems. In this concept the mean square error can be decomposed into a bias term, and a variance term, as follows:

\[ E = B^2 + V \]  

(4)

where \( E \), \( B \), and \( V \) represent respectively the mean square error, the bias, and the variance. The bias represents the consistent offset of the forecast, away from the true value. For example, consider a forecasting model with some tuning parameter. If the forecast is always higher by some amount than the true value (even with different realizations of the error terms of the data generation process), then the bias is positive. It is akin to fitting a linear function to a parabola (with added noise). There will always be that offset from the true value, and this is the bias. The variance represents the variation of the forecast around its mean. So if a forecasting model produces highly variable forecasts for different realization of the error terms, then it has a high variance. Simpler models tend to produce large bias and small variance. On the other hand, complex models produce small bias and large variance. An analysis of the role of bias and variance in forecasting can be found in the work of Ben Taieb and Atiya [3].

As it turns out forecast combination tends to keep the bias little changed or possibly improved. On the other hand, it generally decreases the variance considerably. This is shown as follows. Consider \( y \) to be the variable to be forecasted, and let \( u_1, \ldots, u_N \) be the different forecasts to be combined. Let the forecast combination be

\[ u = \sum_{i=1}^{N} w_i u_i, \quad \text{with} \quad 0 \leq w_i \leq 1, \quad \sum_{i=1}^{N} w_i = 1 \]  

(5)

The bias \( B \) of the combined forecast is given by

\[ B = E(u) - E(y) \]  

(6)

\[ = \sum_{i=1}^{N} w_i E(u_i) - E(y) \]  

(7)
\[
\sum_{i=1}^{N} w_i (E(u_i) - E(y)) \quad (8)
\]
\[
= \sum_{i=1}^{N} w_i B_i \quad (9)
\]
where \(B_i\) is the bias for forecast \(i\), and where the expectation is over the variations of the error terms of the data generation process. So the bias of the forecast combination is the weighted average of the individual biases, and if these biases are comparable, so will be the bias of the combination. On the other hand, Hendry and Clements [6] make the point that it is unexpected to have the biases of the individual forecasts on one side only, so we typically have some cancelations that would reduce the bias. They made an exception during the aftermath of a structural break, where biases would mostly all be on one side, as most models would “stay behind” in a similar way.

Now we consider the variance. Let \(\sigma_i^2\) be the variance of the individual forecast \(i\). Let \(\rho_{ij}\) be the correlation coefficient of forecasts \(i\) and \(j\). The variance is given as follows.

\[
V = E(u - E(u))^2 \quad (10)
\]
\[
= E\left[ \sum_{i=1}^{N} w_i (u_i - \bar{u}_i) \right]^2 \quad (11)
\]
\[
= \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j E[(u_i - \bar{u}_i)(u_j - \bar{u}_j)] \quad (12)
\]
\[
= \sum_{i=1}^{N} w_i^2 \sigma_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j \rho_{ij} \sigma_i \sigma_j \quad (13)
\]
\[
= \left( \sum_{i=1}^{N} w_i \sigma_i \right)^2 - 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j (1 - \rho_{ij}) \sigma_i \sigma_j \quad (14)
\]

By the RMS-arithmetic weighted mean inequality, we can write:

\[
\sum_{i=1}^{N} w_i \sigma_i \leq \sqrt{\sum_{i=1}^{N} w_i \sigma_i^2} \quad (15)
\]

Hence

\[
V \leq \sum_{i=1}^{N} w_i \sigma_i^2 - 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j (1 - \rho_{ij}) \sigma_i \sigma_j \quad (16)
\]
The first term is the average of the variances of the individual forecasts. The second term is positive since \( \rho_{ij} \leq 1 \). One can therefore observe that a substantial positive term (consisting of many more terms) is subtracted from the first term (the average variance), indicating that the variance of the forecast combination tends to decrease considerably. The degree of decrease in variance becomes larger if the correlation coefficients among the constituent forecasts are smaller, confirming the aforementioned finding that diversity improves the performance of the forecast combination. One has to concede though that in most cases the correlation coefficients of the individual forecasts is positive (for example it is typically larger than 0.5). The reason is that ultimately they are forecasting the same quantity, so they are all moved together by underlying fluctuations of the data generation process.

5 Conclusion

In this paper we provided some analysis on why forecast combinations are successful. It is a short peek into some aspects of forecast combination. There are many other very insightful works in the literature into this important topic. They consider several different aspects, such as the effect of serial correlation, heteroscedasticity, structural breaks, estimation error in combination weights, etc. We hope that our work will be one extra step in guiding researchers and practitioners towards successful use and understanding of forecast combinations.

References


