International Journal of Forecasting 31 (2015) 943–951

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems

Andreas Graefe [a,*], Helmut Küchenhoff [b], Veronika Stierle [b], Bernhard Riedl [b]

[a] *Department of Communication Science and Media Research, LMU Munich, Germany*

[b] *Institute of Statistics, LMU Munich, Germany*

ARTICLE INFO

Keywords:
Bayesian analysis
Combining forecasts
Economic forecasting
Election forecasting
Equal weights

ABSTRACT

We compare the accuracies of simple unweighted averages and Ensemble Bayesian Model Averaging (EBMA) for combining forecasts in the social sciences. A review of prior studies from the domain of economic forecasting finds that the simple average was more accurate than EBMA in four studies out of five. On average, the error of EBMA was 5% higher than that of the simple average. A reanalysis and extension of a published study provides further evidence for US presidential election forecasting. The error of EBMA was 33% higher than the corresponding error of the simple average. Simple averages are easy both to describe and to understand, and thus are easy to use. In addition, simple averages provide accurate forecasts in many settings. Researchers who are developing new approaches to combining forecasts need to compare the accuracy of their method to this widely established benchmark. Forecasting practitioners should favor simple averages over more complex methods unless there is strong evidence in support of differential weights.

1. Introduction

Ensemble Bayesian Model Averaging (EBMA) is a relatively new approach to combining forecasts that emerged from the data-heavy domain of weather forecasting (Raftery, Gneiting, Balabdaoui, & Polakowski, 2005). EBMA calculates a weighted average of forecasts, where the weights are based on the past performance and uniqueness of each component forecast. Montgomery, Hollenbach, and Ward (2012), hereafter MHW, test the performance of EBMA for three subject areas (insurgencies, US presidential elections, and US Supreme Court decisions) within the domain of political forecasting. MHW find that the EBMA forecasts are more accurate than the individual component forecasts in each case and propose that the method be used widely for forecasting social science problems.

Combined forecasts are often more accurate than the individual component forecasts. In a meta-analysis of thirty studies, the simple unweighted average of multiple forecasts had errors 12% lower than those of the typical individual forecast (Armstrong, 2001). In addition, the average forecasts were often more accurate than the most accurate component forecast. Armstrong's analysis also indicated that the gains in accuracy from combining are expected to be highest when five or more forecasts can be obtained, when the forecasts draw upon different validated methods and data, and when there is uncertainty as to which forecast is most accurate.

Graefe, Armstrong, Jones, and Cuzán (2014) analyzed the gains from combining forecasts when forecasting US presidential elections by averaging forecasts within and across four different methods: polls, prediction markets, expert judgment, and quantitative models. This approach, the results of which are published at PollyVote.com, yielded large gains in accuracy, much larger than the 12%

* Corresponding author.
E-mail address: a.graefe@lmu.de (A. Graefe).

http://dx.doi.org/10.1016/j.ijforecast.2014.12.001
0169-2070/

error reduction previously estimated by Armstrong (2001). Across the six elections from 1992 to 2012, the combined PollyVote forecast was more accurate than any one of the component methods; on average, the error reductions ranged from 16% (compared to prediction markets) to 59% (compared to polls).

We applaud MHW for their promotion of the use of combined forecasts. By focusing on combining in their study, they raised awareness of a powerful method that is still underutilized in both research and practice in many fields. In most situations, people will make better predictions by combining forecasts from several sources, rather than by relying on a single source.

However, there is a pitfall. EBMA involves a level of complexity that is often unnecessary when combining forecasts. Innumerable studies on combining have shown that simple combining procedures, such as calculating unweighted averages of forecasts, provide a benchmark that is hard for more complex approaches to beat (Clemen, 1989). MHW do not compare EBMA to this widely accepted benchmark adequately, nor do they discuss the conditions under which EBMA is expected to be useful.

The present study summarizes prior evidence on the relative performances of variants of EBMA and the simple average for combining economic forecasts. We then provide new evidence by reanalyzing and extending MHW's analysis of US presidential election forecasts. We find that EBMA contributes little to the accuracy of simple averages when combining forecasts for social science problems.

2. The issue of weights in combining forecasts

A widespread concern when combining forecasts is the question of how best to weight the components, and many scholars have proposed different methods for doing so. However, an early review of more than two hundred published papers from the fields of forecasting, psychology, statistics, and management concluded that the question of *how* to combine forecasts does not seem to be critical to the forecast accuracy. In fact, it was found that the simple average (i.e., assigning equal weights to components) often provides more accurate forecasts than complex approaches to estimating "optimal" combining procedures (Clemen, 1989).

The empirical research since then has repeatedly confirmed these findings. A recent example is the large-scale comparison of simple averages and various sophisticated approaches to the combination of economic forecasts from the European Central Bank's Survey of Professional Forecasters by (Genre, Kenny, Meyler, & Timmermann, 2013). The sophisticated methods included combinations based on principal components, trimmed means, performance-based weighting, optimal least squared estimates, and Bayesian shrinkage. The performances of these methods varied over time, across target variables, and across time horizons. Moreover, any predictive gains relative to an equal weighting of forecasts were shown to be due probably to chance. The authors therefore concluded that there is only a modest case for considering combinations other than equal weighting as a means of summarizing the survey replies better.

An analysis of the relative performances of several combining procedures based on a seven-country data set of economic forecasts made over the period from 1959 to 1999 provided similar results (Stock & Watson, 2004). Simple averages of all available forecasts provided more accurate predictions than sophisticated combination methods, which relied heavily on historical performances for weighting the component forecasts.

One reason for the strong performance of equal weights is the fact that the accuracy of the component forecasts varies over time and depends strongly on external effects. For example, in the study by Stock and Watson (2004), the accuracies of individual forecasts were influenced heavily by economic shocks and political events. Therefore, a good performance in one year or country did not predict a good performance in another. In such a situation, differential weights are of course of limited value.

Another possible explanation is estimation error in the differential weights. Smith and Wallis (2009) investigated this question by conducting a Monte Carlo simulation of combinations of two forecasts, and reappraising a published study using different combinations of multiple forecasts of US output growth. They concluded that the simple average will be more accurate than estimated "optimal" weights if two conditions are met: (1) the combination is based on a large number of individual forecasts and (2) the optimal weights are close to equality. The reason for this is that, in such situations, each forecast has a small weight, and the simple average provides an efficient trade-off against the error that arises from the estimation of weights.[1]

3. Ensemble Bayesian model averaging

Researchers' quest for an optimal solution to forecast combination continues in many fields. EBMA is a relatively new approach to the differential weighting of forecasts, and has become popular in the data-heavy domain of weather forecasting (Raftery et al., 2005). Simply put, EBMA calculates a probabilistic forecast distribution as a weighted average of component forecasts. The weights of the component forecasts are based on a statistical analysis of each individual component's past performance and the

[1] This body of literature is related closely to that on the relative performances of using equal and differential weights for the predictors in linear models. In many fields, the common procedure when developing linear models for predicting any kind of target variable is to identify a subset of most important predictors and to estimate the weights that provide the best possible solution *for a given sample*. The resulting "optimally" weighted linear composite is then used *when predicting new data*. While this approach is useful in situations with large and reliable datasets and few predictor variables, a large body of empirical and analytical evidence shows that the estimation of variable weights from the data is of little, if any, value in situations with small and noisy datasets and large numbers of predictor variables. Graefe (in press) provides an overview of this literature and demonstrates the gains from weighting the predictors equally in linear models for the task of US presidential election forecasting. Across the ten elections from 1976 to 2012, equally weighting the predictors used by established regression models reduced the forecast error for six of the nine models. On average across the ten elections and nine models, equal weights reduced the error of the original regression models by 5%.