Revisiting the forecast combination puzzle: An empirical study

A research proposal submitted for the degree of

Bachelor of Commerce (Honours)

by

Xiefei Li

30204232

xlii0145@student.monash.edu

Supervisor: David T. Frazier

David.frazier@monash.edu

Department of Econometrics and Business Statistics

Monash University

Australia

April 2023

Contents

Introduction

1.1 Research Objective

This thesis aims to investigate the determinants behind, and evidence for the forecast combination puzzle in various domains, and to empirically examine a general solution to the forecast combination puzzle. The combination puzzle refers to the well-known empirical finding that an equally weighted combination of forecasts generally outperforms more sophisticated combination schemes. This phenomenon is often found in the point combinations but it also appears in the density combinations. In this project, we will work with the density distribution and point forecasts are implicitly included. The use of density forecasts also offers forecasters a more comprehensive view of the target variable. Over the past 50 years, the empirical studies undertaken so far have focused more on different time series settings. Thus, one of the main contributions of this research will be to investigate the presence of the combination puzzle in settings outside of pure time series models with density forecasts. As an additional contribution, we will assess the veracity, and applicability, of a recently proposed solution to the forecast combination puzzle suggested in Zischke et al. (2022) and Frazier et al. (2023).

1.2 Literature Review and Motivation

The accuracy of forecasts is of critical concern for forecasters and decision makers. An idea of combining multiple forecasts from different models was originally proposed in the seminal work of Bates and Granger (1969). With the evidence of dramatic improvements in the forecast accuracy, forecast combinations have attracted increasing attention and contributions in the literature, both theoretical and applied (Clemen, 1989; Timmermann, 2006). In short, forecast combination methods involve producing point or density forecasts and then combining them

based on a rule or weighting scheme. This process can sometimes capture more meaningful characteristics of the true data generating process than using a single model, and allow us to combine the best features of different models within a single framework. Researchers have examined a variety of combination methods for point and density forecasts over the past 50 years, see Wang et al. (2022) for a modern literature review.

In most time series setting under which forecast combinations are employed, a striking empirical phenomenon is often observed, coined by Stock and Watson (2004), as the "forecast combination puzzle". The puzzle is encapsulated by the fact that "theoretically sophisticated weighting schemes should provide more benefits than the sample average from forecast combination, while empirically the simple average has been continuously found to dominate more complicated approaches to combining forecasts" (Wang et al., 2022). In the literature, there are two possible explanations for the puzzle. One concentrates on the estimation uncertainty and instability in combination weight like in Stock and Watson (1998), Stock and Watson (2004) and Smith and Wallis (2009). Another one explores the trade-off between bias and variance as in Elliott (2011) and Claeskens et al. (2016). Recently, Zischke et al. (2022) and Frazier et al. (2023) proposed a new aspect of explanation by investigating the sampling variability in model estimation. They illustrated that, asymptotically, the bias and variability mainly come from the estimation of forecasts.

The goal of this thesis is two-fold: first, to search for empirical evidence of the combination puzzle in settings outside of the usual time series in which it has been found; second, to test the empirical veracity of the theoretical solution to the puzzle found in Frazier et al. (2023), both within, and outside of, the standard time series setting where the puzzle is often observed.

Methodology

The first goal of this paper is to construct linear density forecast combinations with parametric models. The results are anticipated to reveal that forecast combinations can deliver improved accuracy over single models, but are not necessarily superior to forecasts obtained from the equally weighted combination.

The next goal is to estimate the unknown parameters of the constituent models and the weight in a single step, and to compare the accuracy of forecasts based on these combinations against the usual combinations process, as well as the equally weighted combination. To measure differences between these forecasts, we will eventually employ forecast accuracy tests, of the type derived in West (1996), which measure out-of-sample differences between forecasts.

Before explaining the details, the following notations will be used throughout the paper. A vector time series $\mathbf{y}_t$ with a total of T observations will be divided proportionally into two parts, an in-sample period R and an out-of-sample period P. The realization of a target variable y at time t is denoted as y_t. Its future value after the in-sample period is denoted as y_{R+h}, where h is the forecast horizon and $h > 0$. The information set at time t, $\mathcal{F}_t$, is comprised of all observed (and known) realizations of y up to time t, i.e., $\mathcal{F}_t = \{y_1, y_2, .., y_t\}$.

A prediction model M determines the conditional probability density for $\mathbf{y}_t$ with unknown parameters θ_M given the history $\mathcal{F}_{t-1}$, denoted by $f(y_t|\mathcal{F}_{t-1}, \theta_M, M)$. Parameter estimates $\hat{\theta}_M$ are obtained by maximizing the log likelihood function of the conditional probability density for the in-sample period, i.e., $\hat{\theta}_M = \text{argmax} \sum_{t=1}^{R} log f(y_t|\mathcal{F}_{t-1}, M)$. These estimates will be held fixed for the easy of evaluation.

2.1 Forecast Combination Method

Based on the idea of linear pooling (Bates and Granger, 1969; Hall and Mitchell, 2007; Geweke and Amisano, 2011), a linear combination of two predictive densities $f^{(t)}$ is constructed with two constituent predictive densities $f_1^{(t)}$ and $f_2^{(t)}$:

$$f^{(t)}(y) = wf_1^{(t)}(y) + (1-w)f_2^{(t)}(y) \tag{2.1}$$

where h is the future value after the in-sample period (R), and w is the weight allocated to the first model. Through this construction, the sum of two weights is implied to be 1, which is necessary and sufficient for the combination to be a density function (Geweke and Amisano, 2011).

Following the literature on density forecast evaluation, our initial analysis will focus on using log predictive score functions to assess the model estimation and measure the accuracy of our density forecasts; see, e.g., Geweke and Amisano (2011) for a discussion on log score and its use in density forecasting. The log predictive score function of a specific model over the forecast horizon $h = 1, 2, ..., P$ (i.e., the out-of-sample period) is defined as follows:

$$LS = \sum_{h=1}^{P} logf(y_{R+h}|\mathcal{F}_{R+h-1}) \tag{2.2}$$

The weight w assigned to the first model will be estimated by maximizing the log predictive score function over the out-of-sample period:

$$\sum_{h=1}^{P} log\Big[wf_1(y_{R+h}|\mathcal{F}_{R+h-1}) + (1-w)f_2(y_{R+h}|\mathcal{F}_{R+h-1})\Big] \tag{2.3}$$

Preliminary Results

3.1 A Motivating Example

Reconsidering the example in section 3 of Geweke and Amisano (2011), the data used is the daily Standard and Poor's (S&P) 500 index from February 11, 2013 to February 10, 2023 (10 years in total), retrieved via the FRED (2023). We choose $j = 1, \cdots, M$, with $M = 5$ prediction models to study the performance of density predictions across sets of `two-model` pools. Each of the j predictive model has a conditionally Gaussian density, which takes the form $f^{(j)}(y) = f_j(y_t|\mathcal{F}_{t-1}) = N\{y_t; \mu_j, \sigma_j^2\}$, where $N\{x; \mu, \sigma^2\}$ denotes the normal probability density function evaluated at value x with mean μ and variance σ^2. The notation $\mathcal{F}_{t-1}$ denotes all information available at time $t-1$, and we assume that the conditional mean and variance of the models are, up to unknown parameters, known at time $t-1$.

Models include commonly used model types such as autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and linear regression model with ARIMA errors. See Appendix for detailed model specifications.

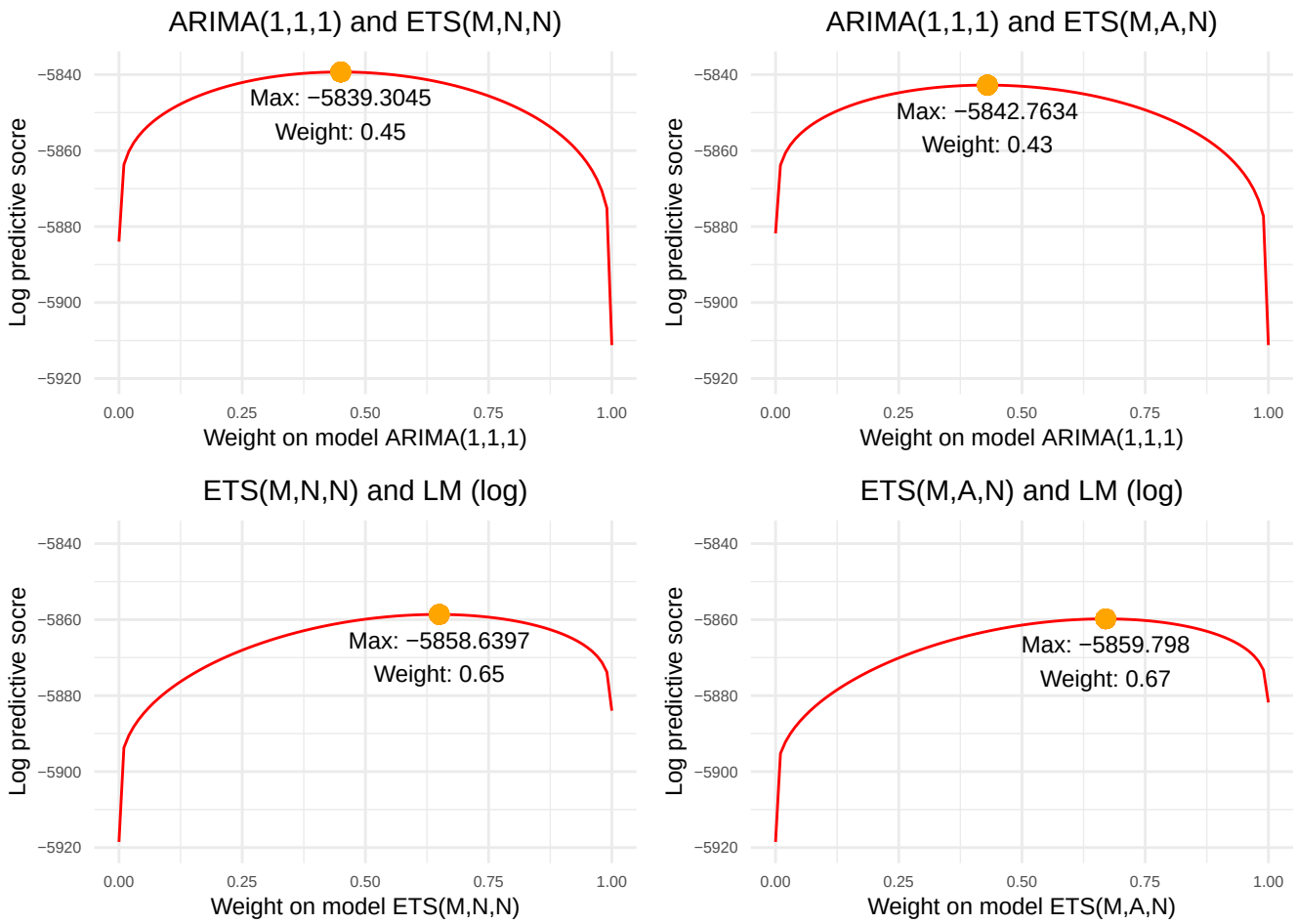
Table 3.1: *Log predictive score of density forecasts combination under two-model pools*

	ARIMA(1,1,1)	ETS(M,N,N)	ETS(M,A,N)	LM (linear)	LM (log)
ARIMA(1,1,1)	*-5911.1974*	-5839.3045	-5842.7634	-5911.1974	-5894.1267
ETS(M,N,N)	0.45	*-5883.9697*	-5881.7790	-5883.9697	-5858.6397
ETS(M,A,N)	0.43	0.08	*-5881.7970*	-5881.7970	-5859.7980
LM (linear)	1	1	1	*-7532.1464*	-5918.5230
LM (log)	0.56	0.65	0.67	0	*-5918.5230*

The diagonal entries contains individual log score.
The log scores for combination pools are located above the diagonal.
Entries below the diagonal show the estimated weight of the model in that column in the two-model pool.

There are 10 sets of two-model combination and each combination log score is generated according to (2.3) with the value of weight changes by 0.01 every time. Table 3.1 shows the estimated weights and combination log scores.

Figure 3.1: *The highest four log predictive scores of weighted two-model-pool combinations for S&P 500 returns predictive densities.*



The weights on the first model is in the x-axis and the corresponding log predictive scores are on the y-axis. Constitutent models are stated in the title. The orange point represent the highest log score of a specific combination. Its value and the corresponding optimal weight are noted below.

Figure 3.1 illustrates the change of the log predictive score for the top 4 combinations as the weight increases. The great improvement in the scores means that the combination forecasts do perform better than the individual forecasts. It is also noticeable that the estimated weights are close to the equal weight (0.5), which could be an evidence for the forecast combination puzzle.

Appendix

The S&P500 returns dataset has a total of 2519 (T) observations and is partitioned into two periods with a rough proportion. The in-sample period contains the first 60% of the data ($R = 1511$), which is used for estimating unknown parameters in each model. The remaining 40% ($P = 1008$) becomes the out-of-sample period for further evaluation.

The following prediction models are used to study the performance of two-model pools:

1. Model 1: An ARIMA(1,1,1) model with an intercept for the natural logarithm of S&P 500.
2. Model 2: An ETS(M,N,N) model for the S&P 500.
3. Model 3: An ETS(M,A,N) model for the S&P 500.

All error terms are assumed to be independent and normally distributed with mean zero and variance σ_j^2 for $j = 1, 2, 3$.

4. Model 4: A linear regression model for the S&P 500 with a trend regressor and errors, follow an ARIMA(1,0,0) process.
5. Model 5: A linear regression model for the natural logarithm of S&P 500 with a trend regressor and errors follow an ARIMA(1,0,0) process.

Both error terms in the ARIMA model are assumed to be independent and normally distributed with mean zero and variance σ_j^2 for $j = 4, 5$.

Exact formulas and explanations of these models can be found in Hyndman and Athanasopoulos (2021). The formula of the conditional variance for the ETS models in this case is discussed in Chapter 6.3 of Hyndman et al. (2008). All coding is performed using R Statistical Software

(version 4.2.1 (2022-06-23)). The packages used are `tidyverse` (Wickham et al., 2019), `dplyr` (Wickham et al., 2023), and `fpp3` (Hyndman, 2023).

All unknown parameters are estimated by maximizing the likelihood function using the in-sample period data. Once the estimated are obtained, they are held fixed for the density evaluation. For each model, we generate the predictive densities at every future time point of S&P 500 returns ($h = 1, 2, ..., P$) given that all past information is known. In order to make a comparison between models, the log of S&P 500 returns will be evaluated by the log normal density function.

The log predictive score of each model is calculated and presented in Table A.1. If only one model can be chosen, the model with the highest score will be preferred, which is the ETS(M,A,N) model with a score of -5881.7970 in this case.

Table A.1: *Log predictive score of each proposed model for S&P 500 returns.*

ARIMA(1,1,1)	ETS(M,N,N)	ETS(M,A,N)	LM (linear)	LM (log)
-5911.1974	-5883.9697	-5881.7970	-7532.1464	-5918.5230

The top 4 combinations and their log scores are collected in Table A.2.

Table A.2: *The top four density forecasts combinations evaluated by the log predictive score*

Combination	Log predictive score
ARIMA(1,1,1) & ETS(M,N,N)	-5839.3045
ARIMA(1,1,1) & ETS(M,A,N)	-5842.7634
ETS(M,N,N) & LM (log)	-5858.6397
ETS(M,A,N) & LM (log)	-5859.7980

Reference

Bates, JM and CW Granger (1969). The combination of forecasts. *Journal of the operational research society* **20**(4), 451–468. DOI: https://doi.org/10.1057/jors.1969.103.

Claeskens, G, JR Magnus, AL Vasnev, and W Wang (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting* **32**(3), 754–762.

Clemen, RT (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting* **5**(4), 559–583.

Elliott, G (2011). Averaging and the optimal combination of forecasts. *University of California, San Diego*.

Frazier, DT, R Zischke, GM Martin, and DS Poskitt (2023). Solving the Forecast Combination Puzzle. [In preparation].

FRED (2023). *S&P500*. https://fred.stlouisfed.org/series/SP500#0 (visited on 02/12/2023).

Geweke, J and G Amisano (2011). Optimal prediction pools. *Journal of Econometrics* **164**(1), 130–141. DOI: https://doi.org/10.1016/j.jeconom.2011.02.017.

Hall, SG and J Mitchell (2007). Combining density forecasts. *International Journal of Forecasting* **23**(1), 1–13.

Hyndman, R and G Athanasopoulos (2021). *Forecasting: principles and practice, 3rd edition*. OTexts.com/fpp3 (visited on 02/12/2023).

Hyndman, R (2023). *fpp3: Data for "Forecasting: Principles and Practice" (3rd Edition)*. R package version 0.5. https://CRAN.R-project.org/package=fpp3.

Hyndman, R, AB Koehler, JK Ord, and RD Snyder (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Smith, J and KF Wallis (2009). A simple explanation of the forecast combination puzzle. *Oxford bulletin of economics and statistics* **71**(3), 331–355.

Stock, JH and MW Watson (1998). *A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series*.

Stock, JH and MW Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of forecasting* **23**(6), 405–430. DOI: https://doi.org/10.1002/for.928.

Timmermann, A (2006). Forecast combinations. *Handbook of economic forecasting* **1**, 135–196.

Wang, X, RJ Hyndman, F Li, and Y Kang (2022). Forecast combinations: an over 50-year review. *International Journal of Forecasting*. DOI: https://doi.org/10.48550/arXiv.2205.04216.

West, KD (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 1067–1084.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686. DOI: 10.21105/joss.01686.

Wickham, H, R François, L Henry, K Müller, and D Vaughan (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.1. https://CRAN.R-project.org/package=dplyr.

Zischke, R, GM Martin, DT Frazier, and DS Poskitt (2022). The Impact of Sampling Variability on Estimated Combinations of Distributional Forecasts. *arXiv preprint arXiv:2206.02376*. DOI: https://doi.org/10.48550/arXiv.2206.02376.