Revisiting the forecast combination puzzle with different data types: 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. 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. 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 probabilistic 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 early forecast comparison studies, Makridakis et al. (1982) and Stock and Watson (1998) demonstrated that combinations perform better than the individual models and a simple average of the forecasts is typically more preferred with empirical evidence. More recently, Stock and Watson (2004), Makridakis, Spiliotis, and Assimakopoulos (2018) and Makridakis, Spiliotis, and Assimakopoulos (2020) delved more into forecast combinations, and also claimed that equally-weighted strategies are highly competitive compared with complicated schemes.

Most explanations for the puzzle concentrate on the error of combination weight estimation. For example, Stock and Watson (2004) showed that the higher average loss and instability make sophisticated weighting schemes to have inferior performance. Timmermann (2006) pointed out that the success of averaging is the efficiency gains under certain assumptions whereas recursive estimation of parameters in the optimally-weighted combinations could potentially lead to biased estimators of the combination weights. On the other hand, Elliott (2011) explored the sizes of theoretical gains from optimal weights and established that the estimation error would outweigh gains if the number of forecasts is small. Later, Claeskens et al. (2016) demonstrated the presence of bias and inefficiency when weights estimation is required, in comparison with the fixed-weights such as the equal weights. While various explanations for the forecast combination puzzle have been suggested over the years (see the above references), a general solution to the puzzle has so far proved elusive.

Instead of exploring the estimation error in determining optimal weights, Zischke et al. (2022) investigated the sampling variability in estimating constituent models and the main determinant of forecast combination performance. They illustrated that, asymptotically, the bias and variability mainly come from the estimation of parameters in the constituent models, rather than the weight estimation. This new insight motivates Frazier et al. (2023) to propose a most general solution to the puzzle. They demonstrated that, in theory, the puzzle is evident due to the way of producing forecast combinations. Furthermore, a process of eliminating the puzzle is also suggested, which is to produce forecasts by estimating parameters and weights simultaneously, if feasible. Under this approach, the sophisticated weighting schemes should (asymptotically) be superior.

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. In addition to point forecasts, the use of density forecasts can offer forecasters or decision markers a broader and more comprehensive view of the target variable (see section 2.6.1. of Petropoulos et al. (2022) for related contributions). 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.

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 $f(y_t|\mathcal{F}_{t-1}, \theta_M, M)$.

2.1 Forecast Combination Method

2.1.1 Log predictive socre function

We use the log predictive score functions to assess individual models and model combinations.

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., $\hat{\theta}_M = argmax \sum_{t=1}^{R} logf(y_t|\mathcal{F}_{t-1}, M)$

For the first step, I will estimate the unknown parameters of each constituent model using Maximum Likelihood Estimation. These estimates will then be held fixed and substituted into their corresponding probability density functions.

Based on the idea of linear pooling (Bates and Granger, 1969; Hall and Mitchell, 2007; Geweke and Amisano, 2011), the linear combinations of two predictive densities $f^{(t)}$ will be 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 $f_1^{(t)}(y)$ and $f_2^{(t)}(y)$ are assumed to follow the normal distributions but with different means and variances, 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).

More specifically, $f_1^{(t)}(y) = f_1(y_t|\mathcal{F}_{t-1}) = N\{y_t; \mu_1, \sigma_1^2\}$ and $f_2^{(t)}(y) = f_2(y_t|\mathcal{F}_{t-1}) = N\{y_t; \mu_2, \sigma_2^2\}$. $N\{x; \mu, \sigma^2\}$ denotes the normal probability density function evaluated at value x with mean μ and variance σ^2. Given $\mathcal{F}_{t-1}$, the conditional mean and conditional variance should be used.

2.2 Evaluation of Models and Weighted Forecast Combinations

This refers to the second step, where I estimate the weight that is assigned to the first model given the aforementioned estimates. The assessment of out-of-sample predictions for individual models and combinations will rely on the average log predictive score function.

The average 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:

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

The optimal weight $w*$ will be estimated by maximizing the average logarithmic predictive score function over the out-of-sample period:

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

The corresponding forecast density combination, given the optimal weight, will be referred to as the optimal combination.

2.3 A Motivating Example

We focused on the combination of two individual forecasts for two reasons, which in most cases apply for the prediction of business figures in enterprises. Typically, a judgmental forecast and one that is derived using purely statistical means are available and corporate planning can be based on one of the forecast or a combination of both forecasts, where additional forecasts cannot be expected to introduce as much additional information. Furthermore, focusing on the two-forecast case allowed us to provide a variety of in-depth analyses. The challenge of extending the model and the decision boundaries to a larger, arbitrary number of fore- casts is subject to future research.

2.3.1 Data

Reconsidering the example in section 3 of Geweke and Amisano (2011), I use 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). Total 2519 (T) available observations are 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.

2.3.2 Model Specification

For simplicity, I use five prediction models 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.

ARIMA is short for autoregressive integrated moving average, and ETS stands for exponential smoothing. All error terms are assumed to be independent and normally distributed with mean zero and variance σ_m^2 for $m = 1, 2, 3$.

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_m^2 for $m = 4, 5$.

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 evaluations. For each model, I 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 each pair of these models, the log of S&P 500 returns will be "back-transformed" by evaluating with the log normal density function.

As a reference, detailed 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).

Preliminary Results

The average log predictive score of each model mentioned in section 2.3.2 is calculated and presented in Table 3.1. 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 -5.8351 in this case. The differences among models seem to be small, disregarding the linear model on the level of S&P 500 returns, but they are closely related to the number of out-of-sample observations and the effect of natural logarithm. Taking these into consideration, the ETS(M,A,N) model could be strongly favored.

Table 3.1: *Average 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)
-5.8643	-5.8373	-5.8351	-7.4724	-5.8716

Besides, there are 10 pairs of two-model combinations given 5 models. For each combination, I generated all the average log predictive scores when the weight on the first model in that combination increases from 0 to 1 by a 0.01 change every time. The optimal combination is generated according to section 2.2. Table 3.2 presents the information about the optimal combination of every pair, including the highest log score and the optimal weight.

Table 3.2: *Average 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)	*-5.8643*	-5.793	-5.7964	-5.8643	-5.8473
ETS(M,N,N)	0.45	*-5.8373*	-5.8351	-5.8373	-5.8121
ETS(M,A,N)	0.43	0.08	*-5.8351*	-5.8351	-5.8133
LM (linear)	1	1	1	*-7.4724*	-5.8716
LM (log)	0.56	0.65	0.67	0	-5.8716

The diagonal entries contains individual average log scores for each model.
The highest average log scores for optimal pools are located above the diagonal.
Entries below the diagonal show the optimal weight of the model in that column in the two-model pool.

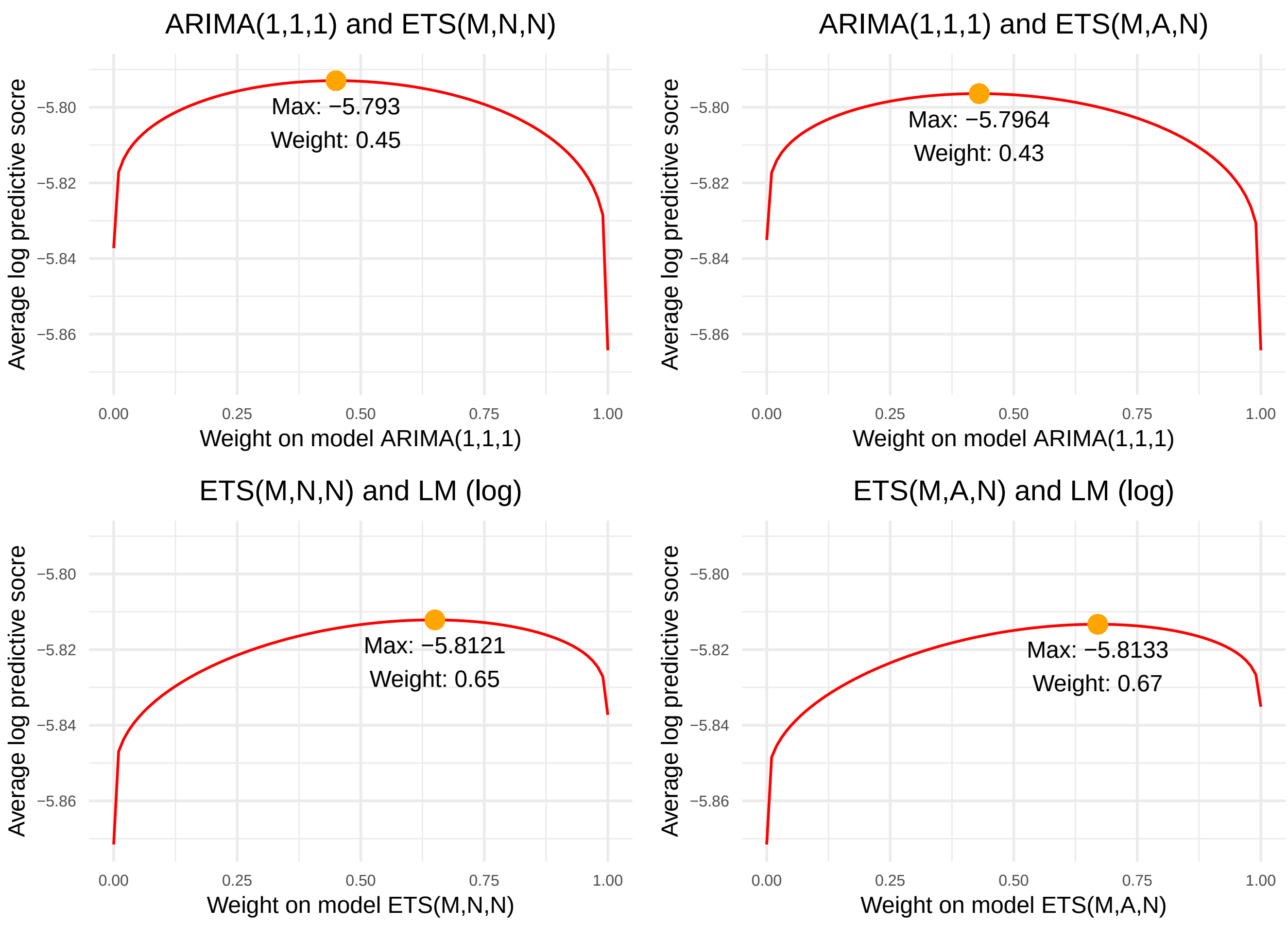
More specifically, I picked the first 4 highest score as shown in Table 3.3.

Table 3.3: *The best four density forecasts combinations evaluated by the average log predictive score*

Combination	Average log predictive score
ARIMA(1,1,1) & ETS(M,N,N)	-5.793
ARIMA(1,1,1) & ETS(M,A,N)	-5.7964
ETS(M,N,N) & LM (log)	-5.8121
ETS(M,A,N) & LM (log)	-5.8133

The Figure 3.1 illustrates the changes in the average log predictive score as the weight increases for the best 4 combinations.

Figure 3.1: *The highest four average 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 average log predictive scores are on the y-axis. Constitutent models are stated in the title. The orange point represent the highest average log score of a specific combination. Its value and the corresponding optimal weight are noted below.

Appendix

All analyses were performed using R Statistical Software (R version 4.2.1 (2022-06-23))

Packages used are `tidyverse` (Wickham et al., 2019), `dplyr` (Wickham et al., 2023), and `fpp3` (Hyndman, 2023).

$$
\begin{aligned}
M_1 : log(y_t) &= \phi_{0,1} + log(y_{t-1}) + \phi_{1,1} log(y)_{t-1} + \theta_{1,1}\epsilon_{t-1} + \epsilon_{t,1} \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_1^2) \\
M_2 : y_t &= \ell_{t-1,2}(1 + \epsilon_{t,2}) \\
\ell_{t,2} &= \ell_{t-1,2}(1 + \alpha_2 \epsilon_{t,2}) \\
M_3 : y_t &= (\ell_{t-1} + b_{t-1})(1 + \epsilon_{t,3}) \\
\ell_t &= \ell_{t-1}(1 + \alpha \epsilon_{t,2}) \\
M_4 : y_t &= \\
M_5 : y_t &=
\end{aligned}
$$

$$
y_t - y_{t-4} = \beta(x_t - x_{t-4}) + \gamma(z_t - z_{t-4}) + \phi_1(y_{t-1} - y_{t-5}) + \Theta_1 \varepsilon_{t-4} + \varepsilon_t \tag{A.1}
$$

Hyndman and Athanasopoulos (2021)

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.

Makridakis, S, A Andersen, R Carbone, R Fildes, M Hibon, R Lewandowski, J Newton, E Parzen, and R Winkler (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting* **1**(2), 111–153.

Makridakis, S, E Spiliotis, and V Assimakopoulos (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting* **34**(4), 802–808.

Makridakis, S, E Spiliotis, and V Assimakopoulos (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **36**(1), 54–74.

Petropoulos, F, D Apiletti, V Assimakopoulos, MZ Babai, DK Barrow, SB Taieb, C Bergmeir, RJ Bessa, J Bijak, JE Boylan, et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*.

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.

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.