# Revisiting the Forecast Combination Puzzle: An Empirical Study

A research thesis 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

August 2023

# Contents

# Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Finish up afterwards.

# Acknowledgements

I would like to express my deepest appreciation to my supervisor David Frazier and my Honours coordinator Heather Anderson for their incredibly valuable guidance, feedback and patience. I am also grateful to receive the Econometrics Honours Memorial Scholarship from Monash University to support my Honours study. Special thanks to my parents, my friends and people who gave me advice and helped me throughout the year.

# Introduction

## 1.1 Research Objective

This thesis aims to investigate the determinants behind, and evidence for the forecast combination puzzle in various domains. The combination puzzle refers to the well-known empirical finding that an equally weighted combination of forecasts generally outperforms more sophisticated combination schemes. While this phenomenon is often referenced in the point forecast combinations literature, it is also present in the literature on density forecast combinations. Starting with two different types of time series datasets, several two-model pools are constructed to explore how the presence of the puzzle is correlated with the in-sample performance of the constituent models.

The empirical studies undertaken so far have focused more on pure time series settings, while there is little literature on the puzzle in the cross-sectional setting. A simulated study is designed to investigate the puzzle in the two-model pool under a regression analysis. In addition, we can derive and obtain a closed-form expression to support findings in the simple regression case. Throughout, we measure the performance of density combinations via the log score function and use mean squared forecast error to assess the accuracy of point combinations.

## 1.2 Literature Review and Motivation

Forecast accuracy is of critical concern for forecasters and decision makers. The application of forecast combination, originally proposed in the seminal work of Bates and Granger (1969), provides the evidence of dramatic improvements in forecast accuracy, and therefore has attracted wide attention and contributions in the literature, both theoretical and applied (Clemen, 1989; Timmermann, 2006). More importantly, this approach often has robust performance across various types of data, proved by numerous empirical results (Geweke and Amisano, 2011). A prominence of researchers also devote efforts on probabilistic forecasting to obtain more information about the uncertainty of the resulting forecast. Similar to point forecasts, researchers now found that density forecast combination outperforms individual density forecast (e.g., Hall and Mitchell, 2007; Geweke and Amisano, 2011).

Forecast combination methods, in general, involve combining multiple forecasts generated from individual or constituent models based on a rule or weighting scheme. Every scheme has its own objective function for producing the "best" forecast combination, along with the optimal weight assigned to each model. This process can sometimes capture more meaningful characteristics of the true data generating process than using a single model, and allows us to combine the best features of different models within a single framework. Researchers have examined a variety of combination methods for both 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 simple 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 other words, complex weighting schemes are designed to improve in-sample accuracy, so these refined forecast combinations should perform better out-of-sample in theory. However, the mean of the contemporaneous forecasts appears to be more robust in practice than forecasts combined through complicated weighting schemes. This finding has been continuously reaffirmed by extensive literature reviews and papers (e.g., Makridakis et al., 1982; Clemen, 1989; Makridakis, Spiliotis, and Assimakopoulos, 2018, 2020), and the simple averaging naturally becomes a benchmark.

The literature explains the puzzle mainly in three aspects: the estimation uncertainty in complicated weighting schemes (Stock and Watson, 1998, 2004; Smith and Wallis, 2009), the bias and inefficiency in the Mean Squared Forecast Error (MSFE) function (Elliott, 2011; Claeskens et al., 2016), and the sampling variability of the forecasts induced via estimation of the constituent model forecasts (Zischke et al., 2022; Frazier et al., 2023). However, all of these explanations implicitly assume that the puzzle will be in evidence when combining forecasts, regardless of the choice of constituent models or the weighing scheme. They ignore the possibility that complicated combination methods can perform much better than the simple average in some cases. In order to make more rigorous explanation statement, we systematically explore the determinants behind the presence of the puzzle under both time series and cross-sectional settings.

Consider a simple case of two-model combination, our initial conjecture is that the presence of the puzzle is highly related to the in-sample fit of two constituent models. When constituent models have similar in-sample fit, the puzzle will be in evidence. Otherwise, the presence of the puzzle is uncertain. Intuitively, the model in-sample performance determines the behavior of forecasts, so forecasts produced by two similarly performed models will not differ much, leading the estimated optimal weight to be around a half. Therefore, it is reasonable to use the simple average method given that the mean of two forecasts is a good estimate and, more importantly, the forecast variance will be halved with no extra parameter estimation. Consequently, we should expect a small difference of the forecast accuracy between the simple averaging and the sophisticated weighting scheme, which is known as the forecast combination puzzle. On the contrary, if two models have distinct in-sample fit, the optimal forecast combination will give more weight to the better one and therefore both weights will be far away from a half. The presence of the puzzle now becomes ambiguous because there are two possible situations in this case, one is that the simple average forecast perform much better than the optimal combination forecast, and the other one is the opposite. By definition, the puzzle is apparently evident in the first case and hard to argue in the second case. Table 1.1 summarizes the hypothesis. We evaluate two constituent models based on their in-sample performance in a relative sense, and also allow models to perform equivalently Bad for different reasons.

| | | $M_2$ | |
|---|---|---|---|
| | | Good | Bad |
| $M_1$ | Good | $\surd$ | ? |
| | Bad | ? | $\surd$ |

**Table 1.1:** *The first row and the first column refer to two constituent models in a combination, $M_1$ and $M_2$. “Good“ means that the model fits the data well, whereas “Bad“ denotes that the model fails to capture some important features of the data. The “$\surd$“ indicates the presense of the forecast combination puzzle, while “?“ implies that the presense of the puzzle is uncertain.*

Even though there is a widespread literature among different pure time series settings, no attention appears to have been given to the cross-sectional setting. We investigate the forecasting performance of two-model pools for simulated cross-sectional data in using simple linear regression models. We derive the mathematical relationship between the optimal weight and elements in the true data generating process (DGP) under the mean squared forecast error. It directly illustrates the determinants of the estimated optimal weight and gives a formal reasoning to interpret empirical findings. Consequently, we find evidence that the forecast combination puzzle is not just about the fit of constituent models but the interaction of the model with the true DGP. In addition, this study provides sufficient empirical evidence to examine the conjecture in Table 1.1 comprehensively. We demonstrate that the logic behind point

The goal of this thesis is two-fold: first, to substantiate the presence of the combination puzzle in the usual time series in which it has been found; second, to explore the relationship between the puzzle and the in-sample fit of constituent models; third, to search for empirical evidence of the combination puzzle in cross-sectional settings; fourth, 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

We will empirically investigate the effect of in-sample performance between two constituent models on the presence of the forecast combination puzzle. Two main scenarios are considered: a small difference of the in-sample accuracy between the models in the pool and a big difference between their in-sample accuracy. The comparison will mainly rely on the log likelihood or the $R^2$, which will be explained shortly.

Next 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.

In the literature, there are several definitions of combinations. We focus on the combination of forecasts from non-nested models for a given dataset, which is commonly performed in two stages:

1. producing separate point or probabilistic forecasts for the next time point using observed data and constituent models, and
2. combining forecasts based on one of the accuracy criteria.

Specifically, we only consider the combination of two individual forecasts, which allows us to delve into interesting and unexplained findings through fast data manipulation.

Before explaining further details, the following notation will be used throughout the paper. An observed time series $y_t$ 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 values 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 parametric model $M$ determines the conditional probability density for $y_t$, denoted by $f(y_t|\mathcal{F}_{t-1}, \theta_M, M)$, given unknown parameters $\theta_M$ and all the past information $\mathcal{F}_{t-1}$. The choice and specification of constituent models vary by the features of the in-sample data. For each model, the error term is assumed to be independent and normally distributed so that the Maximum Likelihood Estimation (MLE) method can be applied to generate the estimators of unknown parameters, i.e., $\hat{\theta}_M = \underset{\theta_M}{\arg\max} \sum_{t=1}^{R} logf(y_t|\mathcal{F}_{t-1}, M)$. Given the log likelihood function of in-sample period for each model, the corresponding estimates are obtained when they maximize that function and then held fixed for out-of-sample procedures. The optimal combination is then constructed with the estimated weight of each model that delivers the best in-sample accuracy.

## 2.1 Density combinations

### 2.1.1 Linear pooling

Consider the case of only two competing models, which we identify through their probability densities. Undoubtedly, densities can be combined in many ways; see Section 3 of Wang et al. (2022) for many popular means of probabilistic combination. One of the commonly used approaches is the "linear opinion pool", which aggregates constituent weighted densities in a linear form (e.g., Bates and Granger, 1969; Hall and Mitchell, 2007; Geweke and Amisano, 2011). For `two-model` pools, constituent densities $f_1(y_t)$ and $f_2(y_t)$ are combined as follows:

$$f(y_t) = \omega\ f_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M1}, M_1) + (1-\omega) f_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M2}, M_2) \tag{2.1}$$

where $\omega$ is the non-negative weight allocated to the probability density derived from the first model. Through this construction, the sum of the model weights is fixed at 1, which is a necessary and sufficient condition for $f(y_t)$ to be a proper density function (Geweke and Amisano, 2011). In addition to producing point forecasts, density forecasts can offer forecasters or decision markers a comprehensive view of the target variable (see section 2.6.1. of Petropoulos et al. (2022) for related contributions).

### 2.1.2 Log socring rules

Following the literature on density evaluation, our initial analysis will focus on using the log score to 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. For each individual model $M$, the log score over the sample $t = 1, \ldots, T$ is:

$$LS = \sum_{t=1}^{T} log\, f(y_t|\mathcal{F}_{t-1}, \hat{\theta}_M, M). \tag{2.2}$$

The "optimal" linear combination is identified to produce the most accurate forecasts when the set of weights maximizes the log score function of two densities over the in-sample observations $y_t, t = 1, 2, \ldots, R,$

$$\hat{\omega}_{\text{opt}} = \arg\max_{\omega} \sum_{t=1}^{R} log\Big[\omega\, f_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M1}, M_1) + (1-\omega)\, f_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M2}, M_2)\Big]. \tag{2.3}$$

Thus, the log predictive score over the out-of-sample period $t = R+1, R+2, \ldots, T$ is:

$$LPS = \sum_{t=R+1}^{T} log\Big[\hat{\omega}_{\text{opt}}\, f_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M1}, M_1) + (1-\hat{\omega}_{\text{opt}})\, f_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_{M2}, M_2)\Big]. \tag{2.4}$$

## 2.2 Point combinations

Although our main focus is the density forecast combination, to simplify certain analysis, point forecast combination is also used. The point forecast of each model corresponds to the mean value of the predicted density distribution. We will use the mean squared forecast error (MSFE), following Bates and Granger (1969) and Smith and Wallis (2009), to measure the accuracy of point forecast combinations in the two-model pools.

### 2.2.1 Linear combination

Similar to the density case, point predictions from two constituent models, $\hat{y}_{1t}$ and $\hat{y}_{2t}$, are aggregated linearly:

$$\hat{y}_t = w\, \hat{y}_{1t} + (1-w)\, \hat{y}_{2t} \tag{2.5}$$

where $\omega$ is the non-negative weight allocated to the point prediction generated from the first model.

### 2.2.2 Mean squared forecast error

The MSFE of an individual model is the average squared difference between the actual value, $y_t$, and the predicted value, $\hat{y}_t$, at each time point over the in-sample period $R$:

$$MSFE = \frac{1}{R}\sum_{t=1}^{R}(y_t - \hat{y}_t)^2. \tag{2.6}$$

The lower the MSFE, the higher the accuracy of the forecast. Therefore, the "optimal" set of weights satisfies that it minimizes the MSFE of the point forecast combination among all other possible sets over the training period:

$$\hat{\omega}_{\text{opt}} = \arg\min_{\omega} \frac{1}{R}\sum_{t=1}^{R} \omega\, \hat{y}_{1t} + (1-\omega)\, \hat{y}_{2t}. \tag{2.7}$$

Consequently, the MSFE over the out-of-sample period $t = R+1, R+2, \ldots, T$ is:

$$MSFE = \frac{1}{P}\sum_{t=R+1}^{T} \hat{\omega}_{\text{opt}}\, \hat{y}_{1t} + (1-\hat{\omega}_{\text{opt}})\, \hat{y}_{2t}. \tag{2.8}$$

## 2.3 Goodness-of-fit

One well-known way of quantifying the fit of a classical linear regression model is its coefficient of determination (R-squared or $R^2$). $R^2$ represents the proportion of explained variation and is often interpreted as the sample variation of the dependent variable explained by regressors in the model. Details and formulas are elaborated in Chapter 2-3c of Wooldridge (2015).

For example, it is known that the in-sample fit of a linear regression model can be represented by $R^2$, which is therefore a suitable measure to determine `Good` and `Bad` models in this context.

# Empirical Results

## 3.1 Pure time series setting (S&P 500)

Reconsidering the example in Section 3 of Geweke and Amisano (2011), the data we use is the daily Standard and Poor's (S&P) 500 index from February 11, 2013 to February 10, 2023 (10 years in total), retrieved from Federal Reserve Economic Data (FRED, 2023). The S&P 500 index 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 to estimate all unknown parameters, including the optimal weight. The remaining 40% ($P = 1008$) becomes the out-of-sample period to evaluate the forecast performance.

We will investigate the presence of the forecast combination puzzle when both models fit the training set well and when one of the model badly fit the data. Constituent models are based on common classes of linear time series models: autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and linear regression model with ARIMA errors. Detailed model specifications for each case will be elaborated in the Appendix.

We choose three predictive models to study the performance of density predictions across sets of `two-model` pools. Each of the $j$ predictive model has a conditional 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 $\mu$ and variance $\sigma^2$. 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 $t-1$.

### 3.1.1 Nonstationary time series

To reduce the level of variability, we take a natural logarithm of the S&P 500 index. Three candidate models are proposed to fit the log of the index, resulting in three sets of two-model combinations in total. The weight $\omega$ will take a value from 0 to 1 and change by 0.01 every time. The log score, as a function of the weight $\omega$, is generated to search for the optimal weight over the in-sample $R$ period (refer to the top row of Figure 3.1). According to equation 2.3, the estimated optimal weight corresponds to the maximum point of the curve. Then we can calculate the log predictive score of the optimal combination for the out-of-sample period based on equation 2.4.

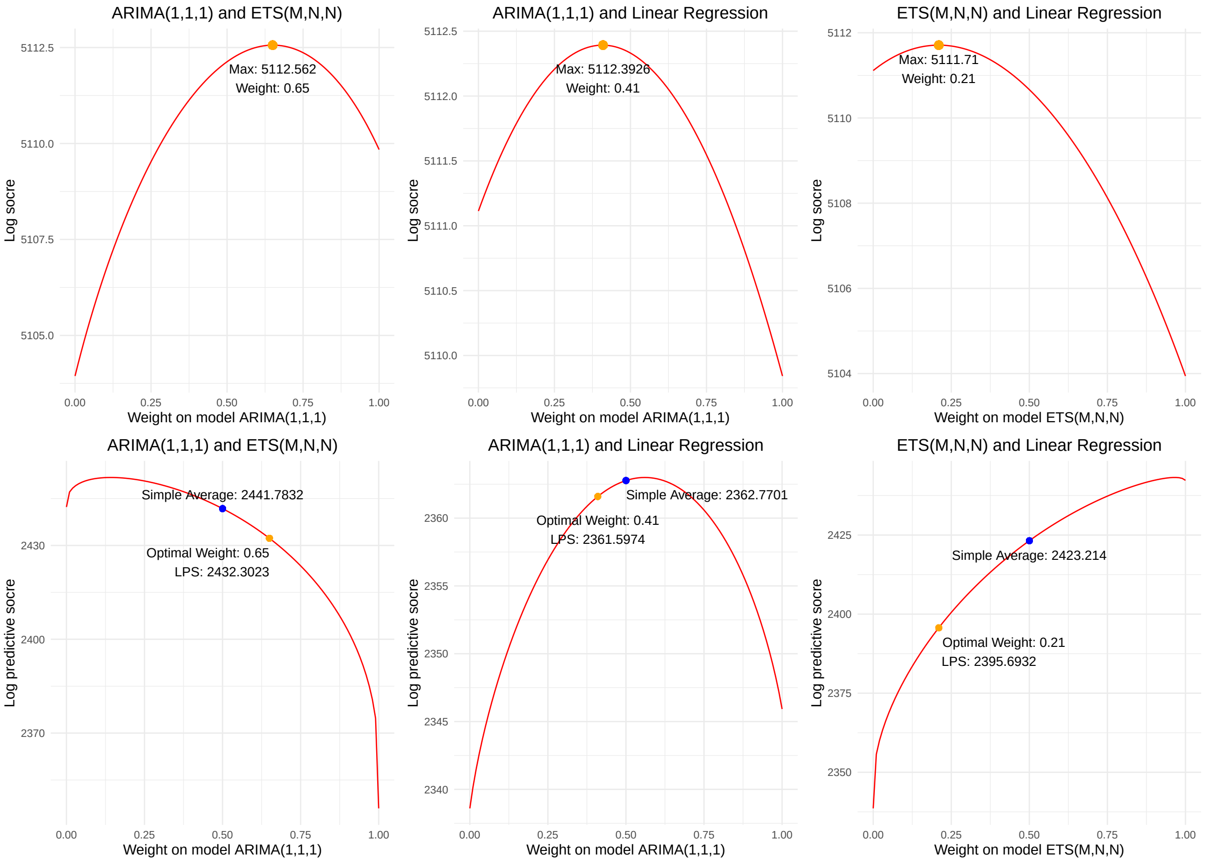

**Figure 3.1:** *Log predictive score of S&P 500 index predictive densities in two-model pools over the in-sample (top) and out-of-sample (bottom) period. Constituent prediction models described in the title. The x-axis represents the weight assigned on the former model of the combination and the y-axis indicates the log predictive score. The orange dot represents the optimal combination, while the blue dot indicates the simple average.*

Figure 3.1 suggests that the forecast combination puzzle is evidenced in all three cases, where the simple average performs better out-of-sample than the optimal combination to different degrees. However, the difference in the log predictive score between two combination methods is varying for each case, and is very likely to be influenced by the in-sample fit of the constituent models. Relevant values are calculated and presented in Table 3.1. It is noticeable that when both models fit the data equally well, i.e., a small difference in the in-sample log score, the prediction accuracy of optimal and simple average approaches is very close. Meanwhile, the optimal forecast combination with a significant gap between the individual in-sample fit ends up performing worse than the mean of forecast densities.

| | Left | Middle | Right |
|---|---|---|---|
| In-sample log score difference | 5.8989 | 1.2718 | 7.1707 |
| Optimal versus average | 9.9569 | 0.2268 | 19.9928 |
| Presence of the puzzle | Yes | Yes | Yes |

**Table 3.1:** *"Left", "Middle" and "Right" correspond to the position of each pair of combination in Figure 3.1. "In-sample log score difference" denotes the absolute difference of the log score over the training period between two models, which can be viewed as an informal accuracy measure of in-sample fit. "Optimal versus average" shows the absolute difference in log predictive score between the optimal combination and the simple average.*

| | Left | Middle | Right |
|---|---|---|---|
| First Model Log Likelihood | 5113.694 | 5113.694 | 1725.137 |
| Second Model Log Likelihood | 1725.137 | 5116.014 | 5116.014 |
| Log Likelihood Difference | 3388.556 | 2.320 | 3390.876 |
| Optimal Weight | 0.65 | 0.41 | 0.21 |
| Puzzle | Yes | Yes | Yes |

**Table 3.2:** *"Log Likelihood Difference" represents the absolute difference of in-sample fit between two models, which is evaluated by the log likelihood. "Optimal Weight" is the estimated weight assigned to the first model in each combination. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

One possible explanation could be that the ETS model fits the training set poorly compared to the other two models, while ARIMA and linear regression perform equally well. Based on the specification of ETS(M,N,N), we may argue that it fails to capture the trend component, shown in Figure 3.2, and is therefore a `Bad` model in the combination. On the other hand, the ARIMA and linear regression can be viewed as `Good` models.

### 3.1.2 Stationary time series

Continuing with the same dataset, we now take a first difference of the log of S&P 500 index, to construct log-returns, and then fit this covariance stationary series. A series is said to be covariance stationary when it has constant mean and variance, and its covariance depends on the time interval only.

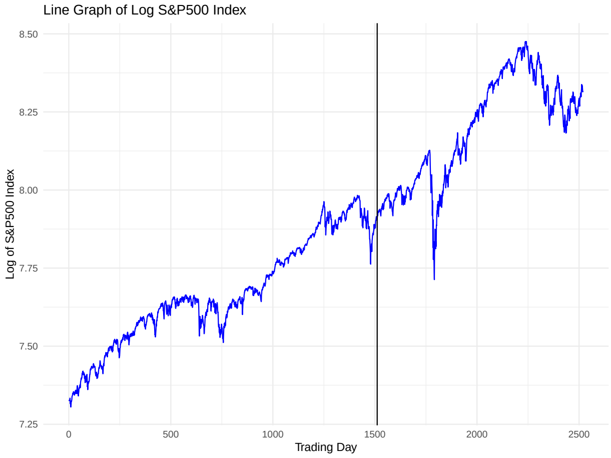

**Figure 3.2:** *The black vertical line separates the traning set and the evaluation set. The training set is on the left and the evaluation set is on the right.*

Consider two candidate models: a Gaussian ARMA(1,1) model and a classical linear regression model with intercept only and ARMA(1,1) errors. To differentiate with the first linear regression model, it is named as Linear Regression 2 in the combination. Figure 3.3 illustrates that two constituent models have a very similar in-sample log score, only 0.0011 difference, and the puzzle is evidenced given only 0.1282 accuracy difference between two forecast combination approaches.

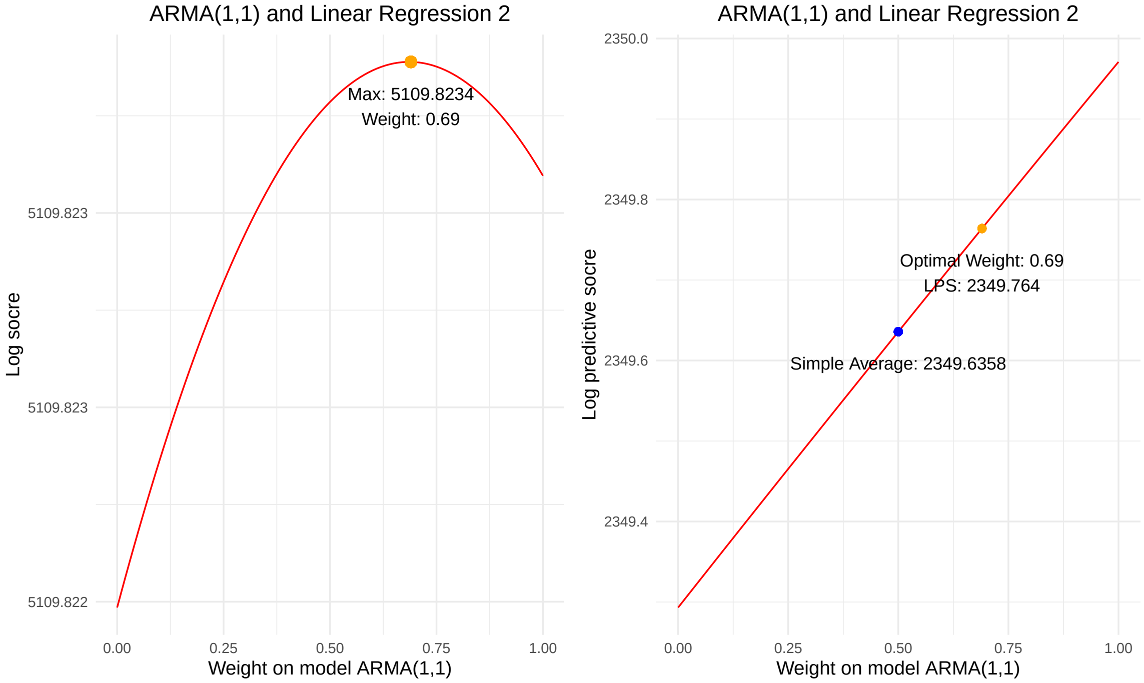

**Figure 3.3:** *Log predictive score of S&P 500 index predictive densities in two-model pools over the in-sample (left) and the out-of-sample (right) period. The x-axis represents the weight assigned on the ARMA(1,1) model and the y-axis indicates the log predictive score. The meanings of colored dots remain the same as before.*

| | ARMA(1,1) and Linear Regression 2 |
|---|---|
| First Model Log Likelihood | 5109.8071 |
| Second Model Log Likelihood | 5109.8054 |
| Log Likelihood Difference | 0.001645751 |
| Optimal Weight | 0.69 |
| Puzzle | Yes |

**Table 3.3:** *"Log Likelihood Difference" represents the absolute difference of in-sample fit between two models, which is evaluated by the log likelihood. "Optimal Weight" is the estimated weight assigned to the first model in each combination. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

This Section 3.1 provides a little empirical evidence for our initial conjecture. When both models fit the data well, i.e., they are `Good` models, then the average density forecast performs almost the same as or slightly better than the optimal density forecast combination, indicating the presence of the forecast combination puzzle. If one model is `Bad` and the other is `Good`, then, at least, the puzzle can be evidenced.

## 3.2 Pure time series with seasonality

With the purpose of further examining our conjecture as to when the puzzle will be in evidence, we now use a quarterly dataset to explore the relationship between the forecast combination puzzle and in-sample model fit. More specifically, we investigate cases where both models are both well-specified (good) or poorly-specified (bad). To simplify the analysis, we produce point forecasts and evaluate point combinations with MSFE.

The data considered is the recorded quarterly total number of unemployed individuals (in thousands) from 1985 Q1 to 2023 Q1, retrieved from the Australia Bureau of Statistics (ABS, 2023). It has a total of 153 ($T$) observations and is slit into two sets in proportion. Same as before, the first 60% of the data ($R = 91$), as the in-sample period, is used to estimate all unknown parameters. The rest 40% ($P = 62$) is the out-of-sample period for the forecast performance evaluation. Also, we use the natural logarithm of the total number of unemployment to reduce the level of variability in the series.

### 3.2.1 Well-specified models

To ensure compatibility with seasonal component, we propose the Seasonal ARIMA (SARIMA) model and the ETS model: ARIMA(2,0,2)(0,1,1)[4] with drift and ETS(A,A,A). The SARIMA is simply an ARIMA model with extra seasonal component. The first parenthesis is same as before. The second parenthesis represents the seasonal AR, integrated, and MA components respectively, separately by the comma. The number in the box bracket indicates the number of observations per year, i.e., the seasonal frequency. An intercept is included in the model. In the ETS model, the seasonal part is reflected by `S` and the third position in the parenthesis. Due to the log transformation, we have additive error, additive trend, and additive seasonality.

The forecast combination puzzle is evidenced when both models are good in Figure 3.4. The optimal forecast point combination has a MSFE of 0.000177 and the simple averaging forecast has a MSFE of 0.000178. The difference between them is negligible. Looking at the in-sample combination plot, two models fit the training set equally well with a difference of 0.00000053. These results exemplify that two `Good` models in a two-model pool will close to having the forecast combination puzzle.

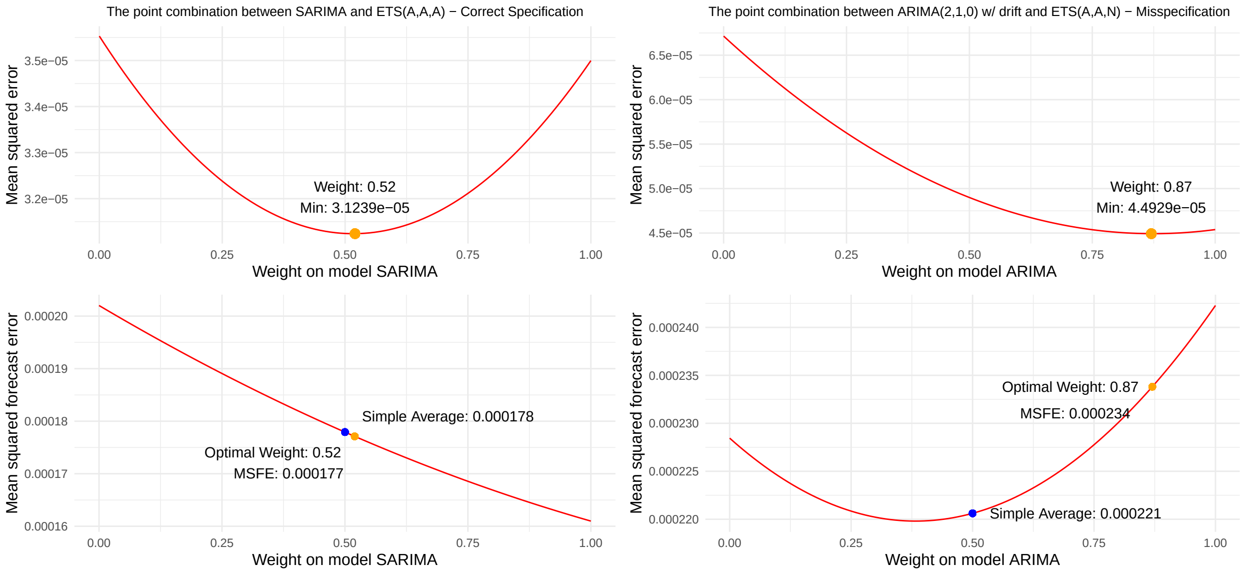

**Figure 3.4:** *MSFE of predictive points in well-specified (left) and pooly-specified (right) two-model pools over the in-sample (top) and out-of-sample (bottom) period. The x-axis represents the weight assigned on the first model and the y-axis indicates the value of MSFE. The meanings of colored dots remain the same.*

### 3.2.2 Poorly-specified models

One way of proposing a `Bad` model for a seasonal dataset is deliberately ignoring the seasonality in the model specification. Even so, we still try to fit the training set well with SARIMA and ETS models but only discarding their seasonal components: ARIMA(2,1,0) with an intercept and ETS(A,A,N).

The right column of Figure 3.4 illustrates that both models have a similar in-sample performance with a deviation of 0.00002175. Furthermore, Figure 3.4 does reveal the forecast combination puzzle, as the simple average performs more superior than the optimal forecast combination with a lower MSFE.

| | Well-specified | Poorly specified |
|---|---|---|
| First Model Log Likelihood | 321.4497 | 322.1642 |
| Second Model Log Likelihood | 260.9102 | 231.9507 |
| Log Likelihood Difference | 60.5395 | 90.2135 |
| Optimal Weight | 0.52 | 0.87 |
| Puzzle | Yes | Yes |

**Table 3.4:** *"Log Likelihood Difference" represents the absolute difference of in-sample fit between two models, which is evaluated by the log likelihood. "Optimal Weight" is the estimated weight assigned to the first model in each combination. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

In this case, we may claim that, regardless whether the constituent models capture all the features of the data, as long as they have similar in-sample fit, the forecast combination puzzle will be evidenced. As a result, even if we have two Bad models, if they have similar in-sample performance, we should expect to find the puzzle.

# Simulation Results

## 4.1 Pure cross-sectional setting

Given that the forecast combination can greatly improve the forecast accuracy, this idea of model combination can also be applied to the cross-sectional setting. A simulated cross-sectional dataset is designed to study how elements in the linear regression model affect the presence of the puzzle, as well as the performance of density combinations. Instead of using real-life data, implementing simulation is easy to control and to make any changes efficiently. At the same time, it is an effective way of validating our conjectures through exploring the forecast combination puzzle from different aspects. In line with previous notations but in the cross-sectional setting, the subscript `t` will change to `i` to represent each individual observation.

### 4.1.1 Experimental design

The true data-generating process (DGP) is assumed to be a linear regression model with only two exogenous and correlated regressors, which satisfies all classical assumptions:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \ \ e_i \overset{i.i.d}{\sim} N(\mu_e, \sigma_e^2) \tag{4.1}$$

where $i$ represents each observation.

The initial set-up has 1000 (N) artificial cross-sectional observations generated from 4.1 with $E[x_{1i}] = E[x_{2i}] = 0$, $Var(x_{1i}) = Var(x_{2i}) = 1$, $Cov(x_{1i}, x_{2i}) = 0.7$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (1, 2, 2)'$, $\mu_e = 0$ and $\sigma_e^2 = 4$.

Following the methodology in Section 2, the data will be divided into an in-sample period (roughly 60%) for estimation and an out-of-sample period for accuracy evaluation. We propose two pooly-specified models to generate density forecasts with each contains only one of the regressors. Assume Model 1 $M_1$ purely includes $x_{1i}$ as the regressor and Model 2 $M_2$ only has $x_{2i}$ as the regressor. The density forecast combinations will follow the construction of `two-model` pools and be evaluated using the log score.

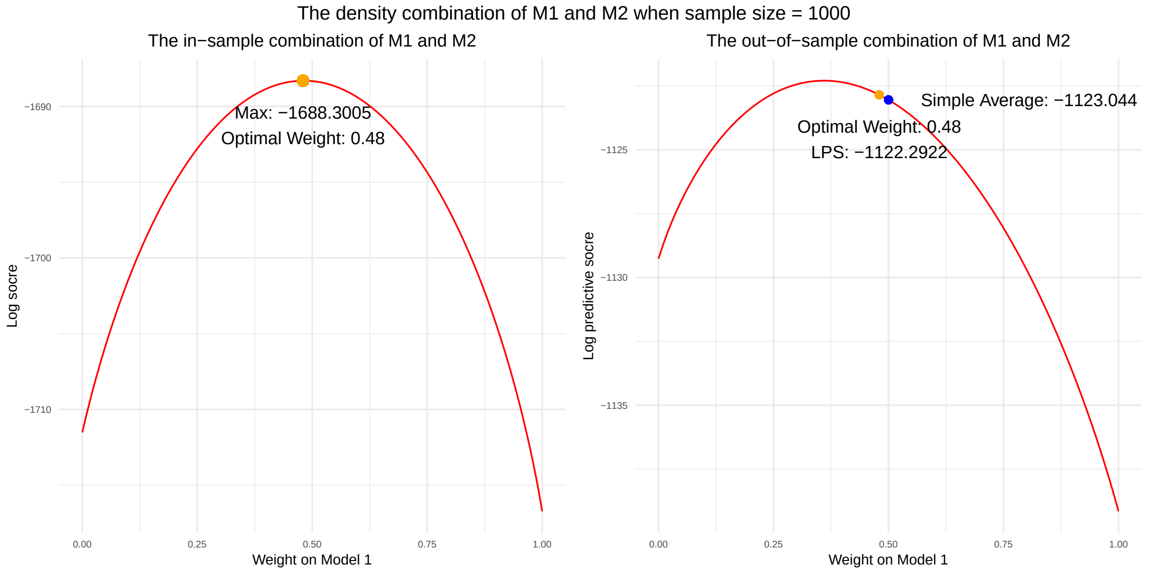

**Figure 4.1:** *Two curves refer to the in-sample (left) and out-of-sample (right) performance of density combinations with artificial cross-sectional data based on the initial set-up. The x-axis represents the weight assigned on Model 1 and the y-axis indicates the log score for each density combination. The orange dot represents the optimal forecast combination, while the blue dot indicates the forecast performance of the simple average combination.*

Figure 4.1 clearly shows that when the sample size is large and two models have similar in-sample performance, the forecast accuracy will be indistinguishable between the simple average of predicted densities and the optimal density forecast combination, which is a strong evidence of the forecast combination puzzle. We can then change the true value of different elements, and determine the conditions under which the puzzle is likely to be evidenced. More rigorously, we will evaluate the in-sample performance with the $R^2$ of constituent models, and then analyse its relationship with the optimal combination weight and the presence of the puzzle.

- **Sample Size**

From Figure 4.2, it is noticeable that the set of optimal weight varies a lot when we have different sample sizes. Model 1 is given an extremely low weight when $N = 50$ whereas it is highly preferred when $N = 100$. The optimal weight is 0.48 when the sample size becomes 1000, shown in Figure 4.1. Based on the log score curve of in-sample combinations, the optimal weight is

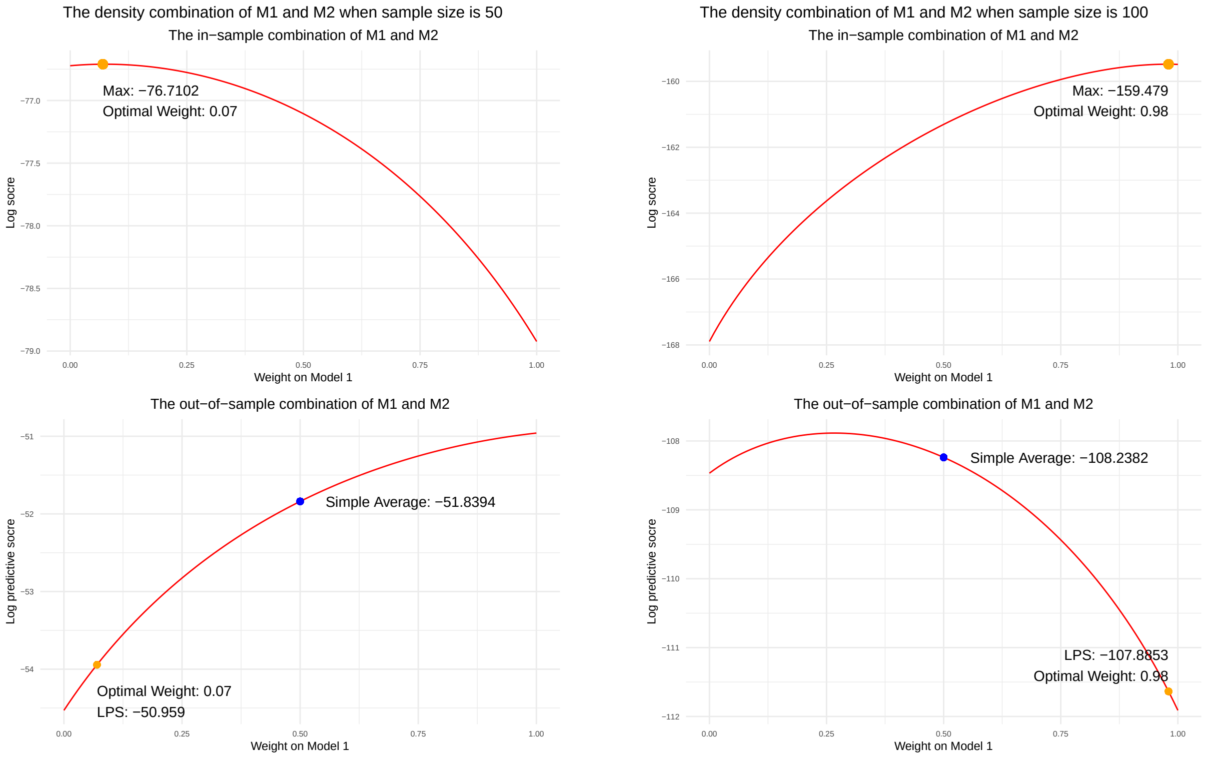

**Figure 4.2:** *Two columns refer to the in-sample combination performance (top) and the out-of-sample combination accuracy (bottom) when $N = 100$ and $N = 1000$ while keeping all others the same as in the initial set-up. The meanings of colored dots are the same as those in Figure 4.1.*

highly correlated with the individual model performance. The number of observations can be viewed as one factor that can affect the model fit. Figure 4.2 also shows that the average density forecast performs much better than the optimal density combination in both cases, i.e., the forecast combination puzzle is found.

| Sample Size | $N = 50$ | $N = 100$ | $N = 1000$ |
|---|---|---|---|
| $R^2$ of $M_1$ | 0.4928 | 0.4953 | 0.3612 |
| $R^2$ of $M_2$ | 0.5620 | 0.3320 | 0.3722 |
| Difference | 0.0692 | 0.1633 | 0.0110 |
| Optimal Weight | 0.07 | 0.98 | 0.48 |
| Puzzle | Yes | Yes | Yes |

**Table 4.1:** *"Difference" represents the absolute difference of in-sample fit between two models. "Optimal Weight" is the estimated weight assigned to $M_1$. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

Table 4.1 illustrates that when two models have a relatively big difference in the in-sample fit $R^2$ (in the second and third columns), we are then more likely to have an extreme optimal weight $\omega$. However, when two models have similar $R^2$ in the fourth column, the optimal weight $\omega$ is close to 0.5. These empirical results first support the conjecture that when models have indifferent in-sample fit, the puzzle is likely evidenced. Additionally, they illustrate that the puzzle can be in evidence when one model performs outstandingly.

- **Magnitude and Sign of $\boldsymbol{\beta}$**

Next, we explore the effect changes in magnitudes or signs of $\beta_1$ and $\beta_2$ given two different sample sizes. From here on, combination plots will be collected and displayed in Appendix A.2. According to Figure A.1, the puzzle is highly sensitive to the absolute difference between two parameters. If the absolute difference is large enough, generally more than half of the smaller coefficient, it is hard to find the puzzle and the optimal combination always wins with a higher log predictive score. In the linear regression analysis, the magnitude of coefficient represents the impact size of corresponding regressor on the dependent variable. A large value of coefficient means that a change in the regressor will affect the dependent variable more in magnitude. Knowing this, it is reasonable to observe that the Model 1 has a decreasing weight in the optimal combination from left to right in Figure A.1. The effect of $x_{2i}$ on $y_i$, $\beta_2$, is relatively larger than the effect of $x_{1i}$ on $y_i$, $\beta_1$, so the Model 2 with $x_{2i}$ should be weighted higher in the combination.

| Different Magnitudes | $\beta_1 = 2,\ \beta_2 = 4$ | $\beta_1 = 2,\ \beta_2 = 6$ | $\beta_1 = 2,\ \beta_2 = 4$ | $\beta_1 = 2,\ \beta_2 = 6$ |
|---|---|---|---|---|
| $R^2$ of $M_1$ | 0.6516 | 0.7057 | 0.4567 | 0.4948 |
| $R^2$ of $M_2$ | 0.6043 | 0.7574 | 0.6082 | 0.7478 |
| Difference | 0.0472 | 0.0517 | 0.1516 | 0.2530 |
| Optimal Weight | 0.59 | 0.32 | 0.18 | 0.04 |
| Puzzle | Yes | No | No | No |
| Sample Size | 100 | 100 | 1000 | 1000 |

**Table 4.2:** *"Difference" represents the absolute difference of in-sample fit between two models. "Optimal Weight" is the estimated weight assigned to $M_1$. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

With reference to the previous results, when the absolute difference is small, the optimal weight $\omega_{opt}$ is expected to be around 0.5 and we are expected to find the puzzle. The second column of Table 4.2 provides another empirical evidence where the absolute difference is around 0.0472. The other three cases, however, illustrate the results when the absolute difference of $R^2$ is big enough. Different from the cases in the second and third columns of Table 4.1, the puzzle is not obvious when one model is more favored, and we have the optimal forecast combination outperforms the simple average forecast. Recall our initial conjecture about the combination of a Good model and a Bad model, simulations have shown some corroborating evidence that the puzzle is ambiguous.

Table 4.3 further justifies our conjecture of the relationship between the in-sample performance and the presence of the puzzle. Especially when the sample size is 100, there is a huge difference between the in-sample fit of two models and $M_2$ is given all the weight in the optimal combination. This clearly implies that the puzzle is not discovered randomly but related to the model in-sample performance. It is also noticeable that conditioning on the same magnitude, the sample size has a large impact on the model fit. When the sample size is small, the absolute difference of in-sample performance becomes larger, leading to an extreme optimal weight and the presence of the puzzle is uncertain as well.

| Different Signs | $\beta_1 = 2,\ \beta_2 = -2$ | $\beta_1 = 4,\ \beta_2 = -4$ | $\beta_1 = 2,\ \beta_2 = -2$ | $\beta_1 = 4,\ \beta_2 = -4$ |
|---|---|---|---|---|
| $R^2$ of $M_1$ | 0.0002 | 0.00002 | 0.0131 | 0.0423 |
| $R^2$ of $M_2$ | 0.1130 | 0.1934 | 0.0321 | 0.0856 |
| Difference | 0.1128 | 0.1934 | 0.0191 | 0.0433 |
| Optimal Weight | 0 | 0 | 0.38 | 0.38 |
| Puzzle | No | No | Yes | Yes |
| Sample Size | 100 | 100 | 1000 | 1000 |

**Table 4.3:** *"Difference" represents the absolute difference of in-sample fit between two models. "Optimal Weight" is the estimated weight assigned to $M_1$. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

- **Variance of regressors**

We keep the variance of $x_{2i}$ the same value and only increase the variance of $x_{1i}$. Then $x_{1i}$ should have a larger variance than $x_{2i}$, thus the variation of $y_i$ can be explained more by Model 1 than Model 2. This can be verified by Table 4.4 where $R^2$ of $M_1$ is always higher than that of $M_2$. Consequently, the in-sample performance difference between the two models is big enough to presume that all four combinations include a good Model 1 and a bad Model 2. As expected in the conjecture, Model 1 should have a higher weight, far away from 0.5, in the optimal combination. Furthermore, the forecast combination puzzle is evidenced in three of them while it is not found in the last situation, indicating that the presence of the puzzle is unclear when there is a big gap in the in-sample fit.

One additional hypothesis is that when the absolute difference of $R^2$ between two models is less than 0.05, they should be treated as having similar in-sample performance. **Formally, the null hypothesis is that the absolute difference of the in-sample fit $R^2$ is less than 0.05.**

| Change in Variance of $x_{1i}$ | $Var(x_{1i}) = 2$ | $Var(x_{1i}) = 4$ | $Var(x_{1i}) = 2$ | $Var(x_{1i}) = 4$ |
|---|---|---|---|---|
| $R^2$ of $M_1$ | 0.5389 | 0.6056 | 0.3981 | 0.4947 |
| $R^2$ of $M_2$ | 0.2899 | 0.2464 | 0.3225 | 0.2536 |
| Difference | 0.2490 | 0.3592 | 0.0756 | 0.2411 |
| Optimal Weight | 0.92 | 0.94 | 0.66 | 0.85 |
| Puzzle | Yes | Yes | Yes | No |
| Sample Size | 100 | 100 | 1000 | 1000 |

**Table 4.4:** *"Difference" represents the absolute difference of in-sample fit between two models. "Optimal Weight" is the estimated weight assigned to $M_1$. "Puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

These results provide a general idea of the relationship between the in-sample fit of constituent models and the presence of the forecast combination puzzle. Based on the new information, the conjecture for two-model pools should be updated, as illustrated in Table 4.5.

| Absolute difference of in-sample fit | Small | Large |
|---|---|---|
| Presence of the puzzle | $\surd$ | ? |

**Table 4.5:** *"Small" means that both models fit the in-sample data equally well (or equally bad), whereas "Large" implies that one of the models performs poorly in fitting the training set. The "$\surd$" implies the presense of the forecast combination puzzle, while "?" means that the presense of the puzzle is ambiguous.*

The choice of model is arbitrary and only the two-model pool is considered. It is also not prudent to determine `small` and `Large` difference based on subjective opinions. Potential improvements include…

# Discussion

# Conclusion

Working in the two-model pools provides an opportunity of exploring a variety of situations in a short period of time. The next challenging step should naturally be to investigate the multiple forecasts combination, and we leave it to future research.

# Appendix

All codes are performed in R Statistical Software (version 4.2.1 (2022-06-23)). The packages used are `tidyverse` (Wickham et al., 2019), `dplyr` (Wickham et al., 2023), `fpp3` (Hyndman, 2023), `gridExtra` (Auguie, 2017), and `mvtnorm` (Genz et al., 2021).

## A.1 Model Specification

The error term, $\epsilon_t$, in each model is assumed to be independent and normally distributed with a zero mean and a constant variance. Each model is independent. Even if using the same notation for unknown parameters across models, the estimators are different.

Exact formulas and explanations of these models can be found in Hyndman and Athanasopoulos (2021). The formula of the conditional variance for the ETS(M,N,N) model is discussed in Chapter 6.3 of Hyndman et al. (2008).

### A.1.1 Nonstationary S&P 500 Index

1. ARIMA(1,1,1) model with an intercept of the natural logarithm of S&P 500 index.

$$log(y_t) = c + log(y_{t-1}) + \phi_1[log(y_{t-1}) - log(y_{t-2})] + \epsilon_t + \theta_1\epsilon_{t-1}$$

2. ETS(M,N,N) model of the natural logarithm of S&P 500 index.

$$\begin{aligned} log(y_t) &= \ell_{t-1}(1+\epsilon_t) \\ \ell_t &= \ell_{t-1}(1+\alpha\epsilon_t) \end{aligned}$$

3. A classical linear regression model of the natural logarithm of the S&P 500 index and ARIMA(1,0,0) errors.

$$
\begin{aligned}
log(y_t) &= \beta_0 + \beta_1 t + u_t \\
u_t &= \phi_1 u_{t-1} + \epsilon_t
\end{aligned}
$$

### A.1.2 Stationary S&P 500 Index

1. ARMA(1,1) model with an intercept of the natural logarithm of S&P 500 returns.

$$
log(y_t) - log(y_{t-1}) = c + \phi_1[log(y_{t-1}) - log(y_{t-2})] + \epsilon_t + \theta_1 \epsilon_{t-1}
$$

2. A classical linear regression model of the natural logarithm of the S&P 500 returns and ARMA(1,1) errors.

$$
\begin{aligned}
log(y_t) &= \beta_0 + u_t \\
u_t &= \phi_1 u_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}
\end{aligned}
$$

### A.1.3 Well-specified models for seasonal unemployment dataset

1. ARIMA(2,0,2)(0,1,1)[4] model with an intercept of the natural logarithm of unemployed individuals.

$$
\begin{aligned}
log(y_t) = c + log(y_{t-4}) + \phi_1[log(y_{t-1}) - log(y_{t-5})] + \phi_2[log(y_{t-2}) - log(y_{t-6})] \\
+ \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Theta_1 \epsilon_{t-4} + \theta_1 \Theta_1 \epsilon_{t-5} + \theta_2 \Theta_1 \epsilon_{t-6}
\end{aligned}
$$

2. ETS(A,A,A) model of the natural logarithm of unemployed individuals.

$$
\begin{aligned}
log(y_t) &= \ell_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha \epsilon_t \\
b_t &= b_{t-1} + \beta \epsilon_t \\
s_t &= s_{t-m} + \gamma \epsilon_t
\end{aligned}
$$

### A.1.4 Poorly-specified models for seasonal unemployment dataset

1. ARIMA(2,1,0) model with an intercept of the natural logarithm of unemployed individuals.

$$log(y_t) = c + log(y_{t-1}) + \phi_1[log(y_{t-1}) - log(y_{t-2})] + \phi_2[log(y_{t-2}) - log(y_{t-3})] + \epsilon_t$$

2. ETS(A,A,N) model of the natural logarithm of unemployed individuals.

$$\begin{aligned} log(y_t) &= \ell_{t-1} + b_{t-1} + \epsilon_t \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha\epsilon_t \\ b_t &= b_{t-1} + \beta\epsilon_t \end{aligned}$$

## A.2 In-sample and out-of-sample combination plot

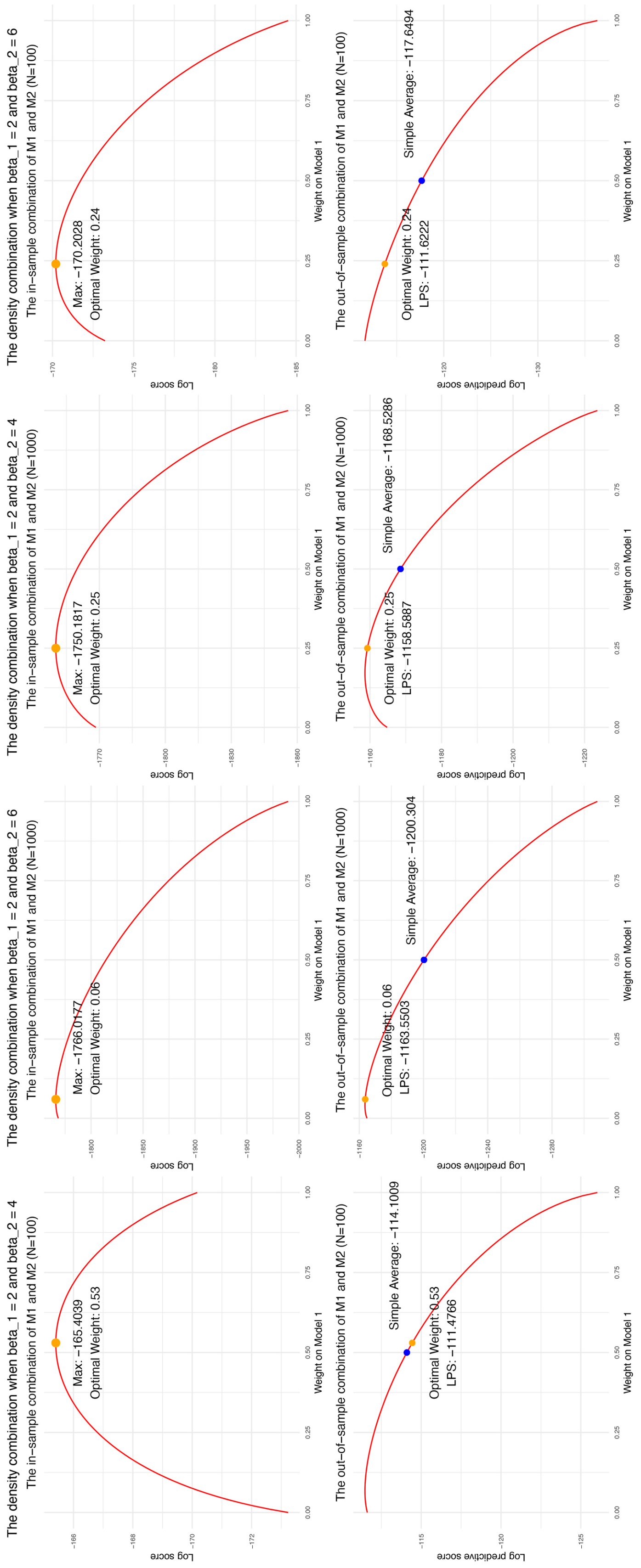

**Figure A.1:** *$\beta_1$ and $\beta_2$ have the same sign but different magnitudes. The first and third columns $\beta_1 = 2$ and $\beta_2 = 4$, and the second and fourth columns $\beta_1 = 2$ and $\beta_2 = 6$. The sample size is indicated in the subtitle. Other variables remain unchanged as in the initial set-up.*

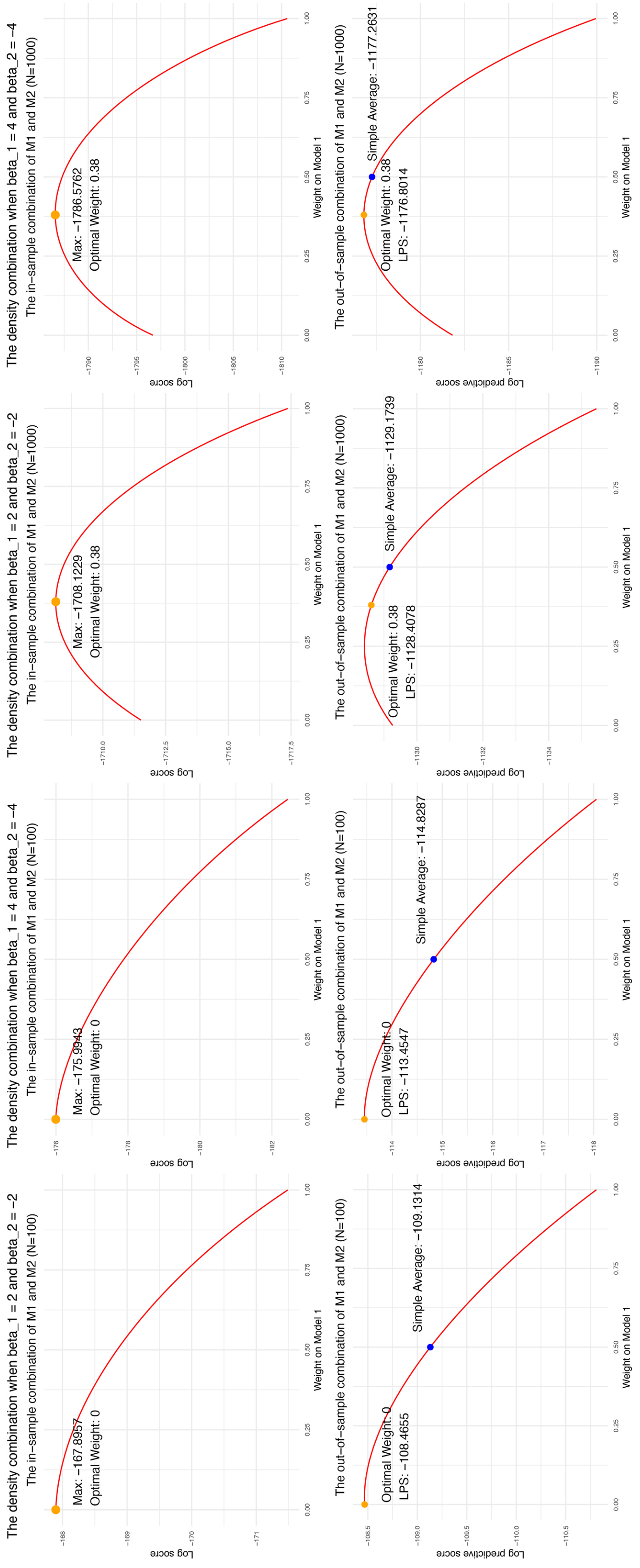

**Figure A.2:** *$\beta_1$ and $\beta_2$ have the same magnitude but different signs, i.e. $\beta_1 = -\beta_2$. The first and third columns $\beta_1 = 2$ and $\beta_2 = -2$, and the second and fourth columns $\beta_1 = 4$ and $\beta_2 = -4$. The sample size is indicated in the subtitle. Other variables remain unchanged as in the initial set-up.*

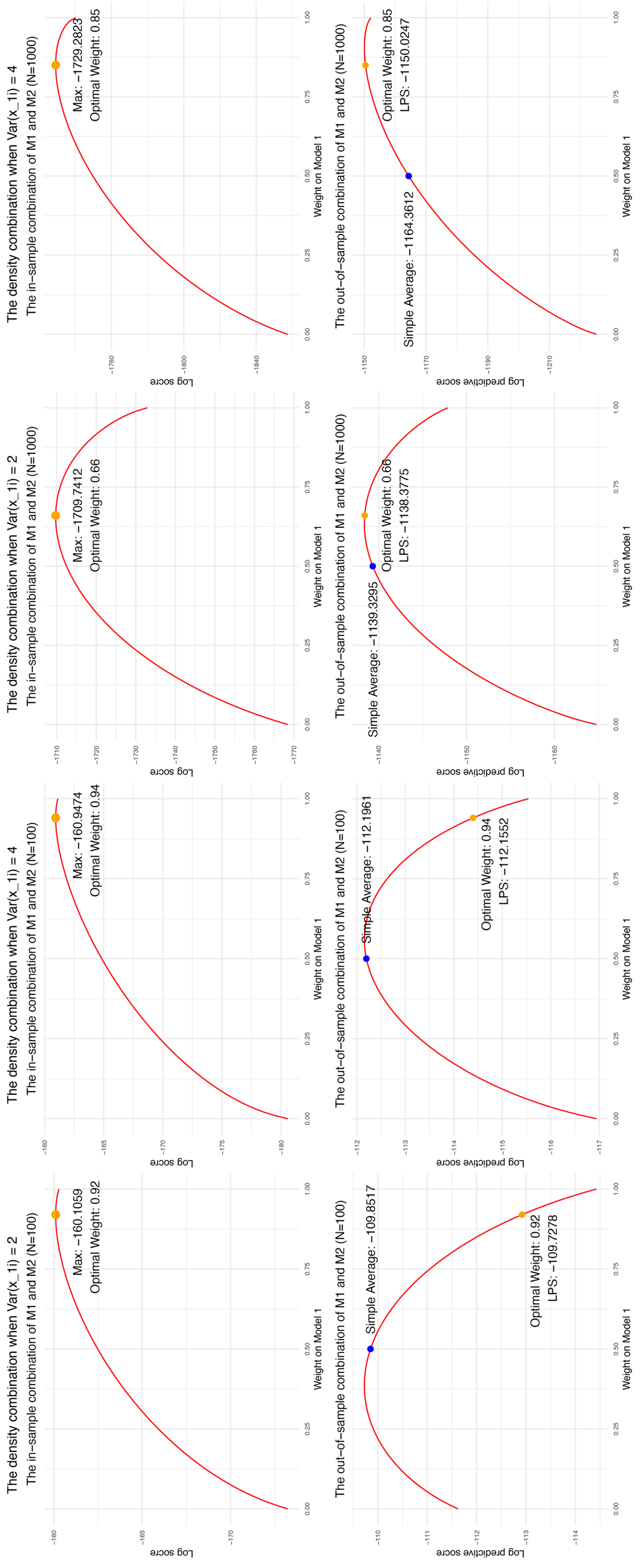

**Figure A.3:** *The first and third columns $Var(x_{1i}) = 2$ and $Var(x_{2i}) = 1$. The second and fourth columns $Var(x_{1i}) = 4$ and $Var(x_{2i}) = 1$. The sample size is indicated in the subtitle. Other variables remain unchanged as in the initial set-up.*

# Reference

ABS (2023). *Labour Force, Australia, Detailed*. https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/latest-release (visited on 03/28/2023).

Auguie, B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

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).

Genz, A, F Bretz, T Miwa, X Mi, F Leisch, F Scheipl, and T Hothorn (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3. http://CRAN.R-project.org/package=mvtnorm.

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*.

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.

Wooldridge, JM (2015). *Introductory econometrics: A modern approach*. Cengage learning.

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.