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

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 thank my supervisor, my Honours coordinator, all the people who helped me throughout the year and myself.

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 forecast combinations but it is also the case in the density forecast combinations. Starting with time series data, this paper explores how the puzzle is affected by the fitness of models on the high volatile or the strong seasonal dataset. The empirical studies undertaken so far have focused more on pure time series settings, while there is little literature on the presence of the combination puzzle in the cross-sectional setting. A simulated study is conducted to study the puzzle in the two-model combination under a regression analysis. The performance of density combinations will be assessed via the log score function and mean squared forecast error is used to assess the point combinations for seasonal data. 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 forecast accuracy is of critical concern for forecasters and decision makers. With the evidence of dramatic improvements in the forecast accuracy, forecast combinations have attracted wide attention and contributions in the literature, both theoretical and applied (Clemen, 1989; Timmermann, 2006). More importantly, this promising approach often has a robust performance for

various types of series, which is borne out by numerous empirical results (Geweke and Amisano, 2011). Makridakis et al. (1982) carefully examined the forecast accuracy with a considerable amount of time series, and reported that forecast combinations perform better than individual models. Later, Stock and Watson (1998) claimed that the best-performing single method can be further improved by incorporating other forecasts, based on empirical comparisons of different forecasting methods. Despite of the point forecasting, researchers also devote efforts on probabilistic forecasting with more information about uncertainties and continue to find that optimal density forecast combination outperforms individual forecasts (e.g., Hall and Mitchell, 2007; Geweke and Amisano, 2011).

Forecast combinations refer to the idea of combining multiple forecasts generated from possible models, which was originally proposed in the seminal work of Bates and Granger (1969). The forecast combination methods, in general, involve producing forecasts from constituent models, and then combining them based on a rule or weighting scheme. Each scheme has different selection criteria for the "best" forecast combination and the corresponding weight value assigned to each model. 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 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 the accuracy, so these refined forecast combinations should perform better in theory. However, the mean of the contemporaneous forecasts appears to be more robust in practice than weighted forecasts combined through complicated schemes. This finding has been reaffirmed by extensive literature reviews and papers (e.g., Clemen, 1989; Stock and Watson, 1998, 2004; Smith and Wallis, 2009; Makridakis, Spiliotis, and Assimakopoulos, 2018, 2020), and simple averaging naturally becomes a benchmark.

There are two possible explanations for the puzzle in the literature. One concentrates on the estimation uncertainty in combination weight (Stock and Watson, 1998, 2004; Smith and Wallis, 2009). Complicated weighting schemes introduce variability and uncertainty when estimating parameters, whereas the simple averaging does not require any estimation. The higher average loss and instability in the study of Stock and Watson (2004) were a strong evidence for the inferior performance of sophisticated weighting schemes. On the other hand, Elliott (2011) and Claeskens et al. (2016) explore the trade-off between bias and variance in the Mean Squared Forecast Error (MSFE). 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. They further proved that equally weighted combination is unbiased and its variance has only one component, resulting in a smaller mean squared error than a biased combination. However, this is applicable and specific to the MSFE scheme.

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. Recently, Zischke et al. (2022) and Frazier et al. (2023) proposed a new explanation for the puzzle in a general way by investigating the sampling variability of the forecasts induced via estimation of the constituent model forecasts (i.e., the models used to produce the forecasts). They illustrated that, asymptotically, the bias and variability mainly come from the estimation of the models used to produce the constituent model forecasts. The common way of producing forecast combinations keeps the model estimation uncertainty fixed during the weight estimation process, which is one reason of having the puzzle. If constituent models and weights can be estimated jointly, if feasible, the puzzle can be eliminated suggested by Frazier et al. (2023). Under this approach, the sophisticated weighting schemes should (asymptotically) be superior.

Following the research direction of Blanc and Setzer (2016), we are looking for a relationship between model fitting and the presence of the puzzle in the case of two-model combination. We speculate that the puzzle is only revealed when both constituent models are correctly specified and fit the data well. If one of them fails to capture the data patterns, the forecast combination will give more weight to the better model. Even if both models are misspecified, the complicated weighing scheme will try to reduce the uncertainty of the worse model as much as possible. Table 1.1 visually summarizes the conjectures.

	M_2	
	Good	Bad
M_1 Good	$\sqrt{}$	$\times$
Bad	$\times$	$\times$

Table 1.1: *The first row and the first column refer to two constituent models in a combination, M_1 and M_2. Good indicates the model is correctly specified and fits the data well, whereas Bad denotes the model is misspecified and fails capture the important features of the data. The "$\sqrt{}$" implies the presense of the forecast combination puzzle, while "$\times$" means no forecast combination puzzle.*

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 cross-sectional data and misspecified models via a simulation study. At a glance, the presence of the puzzle is determined by at least 3 elements in the two-model pools, which are the sample size, the true value of parameters except the intercept, and the variances of regressors.

The goal of this thesis is manifolds: first, to substantiate the presence of the combination puzzle in the usual time series in which it has been found; second, to examine the relationship between the puzzle and the model specification; 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

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.

In the literature, there are several definitions of combinations. We focus on the combination of forecasts from independent 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, considering the combination of two individual forecasts allows us to delve into interesting and unexplained findings through fast data manipulation (Blanc and Setzer, 2016).

Before explaining further details, the following notation 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 values after the in-sample period is denoted as y_{R+h}, where h

7

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., $\mathcal{F}_t = \{y_1, y_2, .., y_t\}$.

The choice and specification of constituent models are determined 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. 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 the following procedures. The optimal combination is then constructed with the estimated weight of each model that delivers the best accuracy.

A parametric model M determines the conditional probability density for $\mathbf{y}_t$, denoted by $f(y_t|\mathcal{F}_{t-1}, \theta_M, M)$, given unknown parameters θ_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., $\hat{\theta}_M = \arg\max_{\theta_M} \sum_{t=1}^{R} \log f(y_t|\mathcal{F}_{t-1}, M)$. 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 the following procedures.

One thing to clarify is that we do not consider the properties of estimators, so model misspecification will not ruin the results. CHECK!!!!

2.1 Density combinations

2.1.1 Linear pooling

Consider the case of only two competing 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": aggregate constituent weighted densities in a linear form (e.g., Bates and Granger, 1969; Hall and Mitchell, 2007; Geweke and Amisano, 2011). For the two-model pools, constituent densities $f_1(y_t)$ and $f_2(y_t)$ are combined as follows:

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

where w is the non-negative weight allocated to the probability density derived from the first model. Through this construction, the sum of two weights is implied to be 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 point forecasts, the use of density forecasts can offer forecasters or decision markers a more 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 function 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 $t = 1, 2, \ldots, R$.

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

Thus, the log predictive score over the forecast horizon $h = 1, 2, \ldots, P$ (i.e., the out-of-sample period) is:

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

2.2 Point combinations

For time series with seasonal patterns, we follow Bates and Granger (1969) and Smith and Wallis (2009), and consider the combination of pairs of point forecasts for the ease of calculation. Same as Smith and Wallis (2009), the mean squared forecast error (MSFE) is adopted to measure the point forecast accuracy.

2.2.1 Linear combination

Similar to the density case, point predictions from two 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 w is the non-negative weight allocated to the point prediction generated from the first model.

2.2.2 Mean squared forecast error (MSFE)

The MSFE is defined as

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

The "optimal" set of weights satisfies that it minimizes the MSFE among all other possible sets.

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 the FRED (2023). The S&P500 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 models: autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and linear regression model with ARIMA errors. Detailed model specifications for each case will be clarified in the corresponding section.

We choose three prediction 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 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 $t-1$.

3.1.1 Nonstationary time series

To reduce the level of variability, we take a natural logarithm of the S&P 500 index y_t and fit the data directly without removing its stochastic trend with three candidate models.

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.

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

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

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

The error term, ϵ_t, in each model is assumed to be independent and normally distributed with a zero mean and a constant variance.

There are three sets of two-model combinations in total. Consider the weight w takes a value from 0 to 1 and changes by 0.01 every time. The log score, as a function of weight, 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 suggests that the forecast combination puzzle is only evident in the second predictive density combination (ARIMA(1,1,1) and Linear Regression). In the middle column, the log predictive score of the optimal combination, 2361.5974, is very close to that of the simple average

Table 3.1: *Density forecasts combination under two-model pools of S&P 500*

	ARIMA(1,1,1)	ETS(M,N,N)	LR
ARIMA(1,1,1)	*2345.9262*	*2441.7832*	2362.7701
ETS(M,N,N)	0.65	*2442.2965*	2423.2140
LR	0.41	0.21	*2338.6005*

The diagonal entries contains individual log predictive score calculated with *R* observations.

The log predictive scores of optimal combinations with *R* observations are located above the diagonal.

Entries below diagonal show the estimated optimal weight of the model in that column.
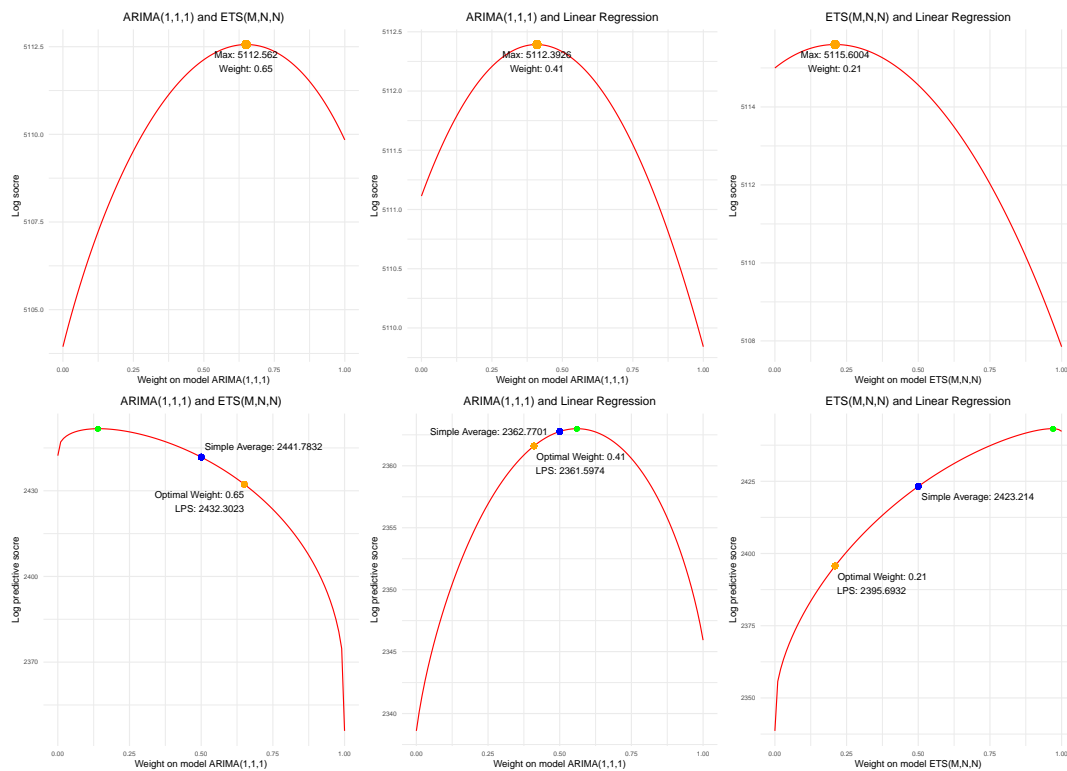

Figure 3.1: *Log predictive score of S&P500 index predictive densities in two-model pools over the in-sample (top) and out-of-sample (bottom) period. Constituent prediction models described in the title. The x-axis represents the weight assigned on the former model of the combination and the y-axis indicates the log predictive score. The orange dot represents the optimal set of weights and the corresponding log predictive score in each case, while the blue dot indicates the forecast performance of the simple averaging method. The green dot, as a reference, refers to the maximum point of the out-of-sample curve.*

combination, 2362.7701. However, the other two combinations do not reveal the puzzle. Besides, the simple average performs superior than the optimal combination in both cases.

WHICH ONE???

ETS too bad?

One possible explanation could be that the ETS model badly fits the training set but forecasts well for the evaluation set in both cases. Given this data split, the optimal weight set gives
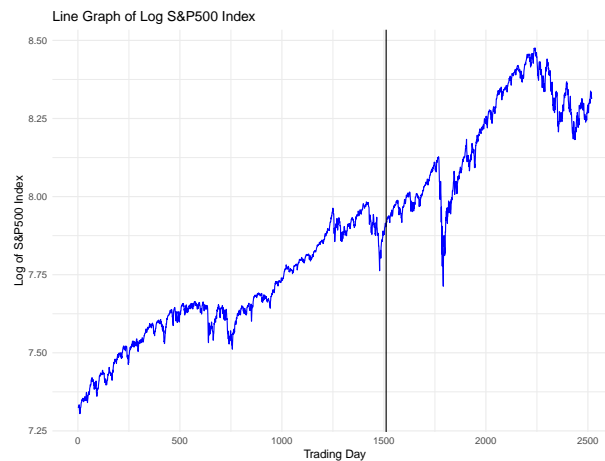
Figure 3.2: *The black vertical line separates the traning set and the evaluation set. The training set is on the left and the evaluation set is on the right.*

less weight on the ETS model, leading to a low log predictive score for the optimal predictive density combination, compared with the simple average. This phenomenon implies that the optimal weight set could not be fixed but a varying variable when one of the constituent models does not fit well with the data. We conjecture that there is a relationship between the forecast combination puzzle and the fitness of constituent models.

ARIMA & LR too good?

Since we are modelling the log of S&P500, its trending behavior and structural breaks are dominant and need to be taken care of when fitting the model. That means a good model should capture the trend, deal with the structural breaks and incorporate all important features of this nonstationary data.

One possible explanation could be that the ARIMA model and the linear regression model fit the training set too well to take any future changes into consideration. It is also unanticipated to incorporate the large structural break and the changing slopes of the trend in the evaluation set as shown in Figure **??**. Thus, their density forecast combinations end up with lower log predictive scores, compared with the simple average.

3.1.2 Stationary time series

Continuing with the same dataset, we now fit the stationary series by taking a first difference of the log of S&P500. A series is said to be stationary when it has constant mean and variance, and its covariance depends on the time interval only. In other words, the entire series should

have a roughly consistent pattern. Then, the training set and the evaluation set will not have completely different behaviors that highly influence the goodness-of-fit of models.

Consider two candidate models: a Gaussian ARMA(1,1) model and a classical linear regression model with ARMA(1,1) errors. Figure 3.3 shows that two constituent models have a very similar in-sample accuracy and the puzzle is obvious in the forecast combination. As expected, since both models fit the data well, the simply averaged point forecast performs almost the same as the optimal combined point forecast, indicating the presence of the forecast combination puzzle.
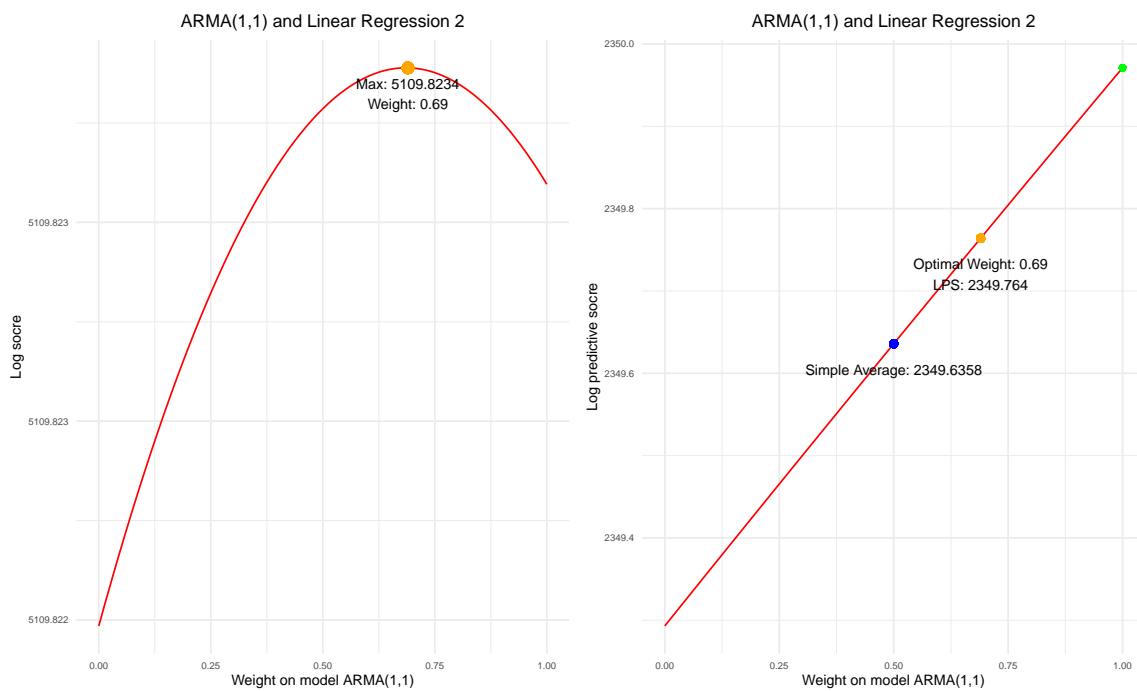
Figure 3.3: *Log predictive score of S&P500 index predictive densities in two-model pools over the in-sample (left) and the out-of-sample (right) period. The x-axis represents the weight assigned on the ARMA(1,1) model and the y-axis indicates the log predictive score. The meanings of colored dots remain the same.*

## 3.2 Pure time series with seasonality

With the purpose of corroborating above conjectures, we experiment with a quarterly dataset to observe the relationship between the forecast combination puzzle and the specification of models. More specifically, we would like to investigate cases when models are both correctly specified or both misspecified. To make our life easier, we produce point forecasts and evaluate point combinations with Mean Squared Forecast Error (MSFE).

The data considered is the 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 seires.

3.2.1 Correctly 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.
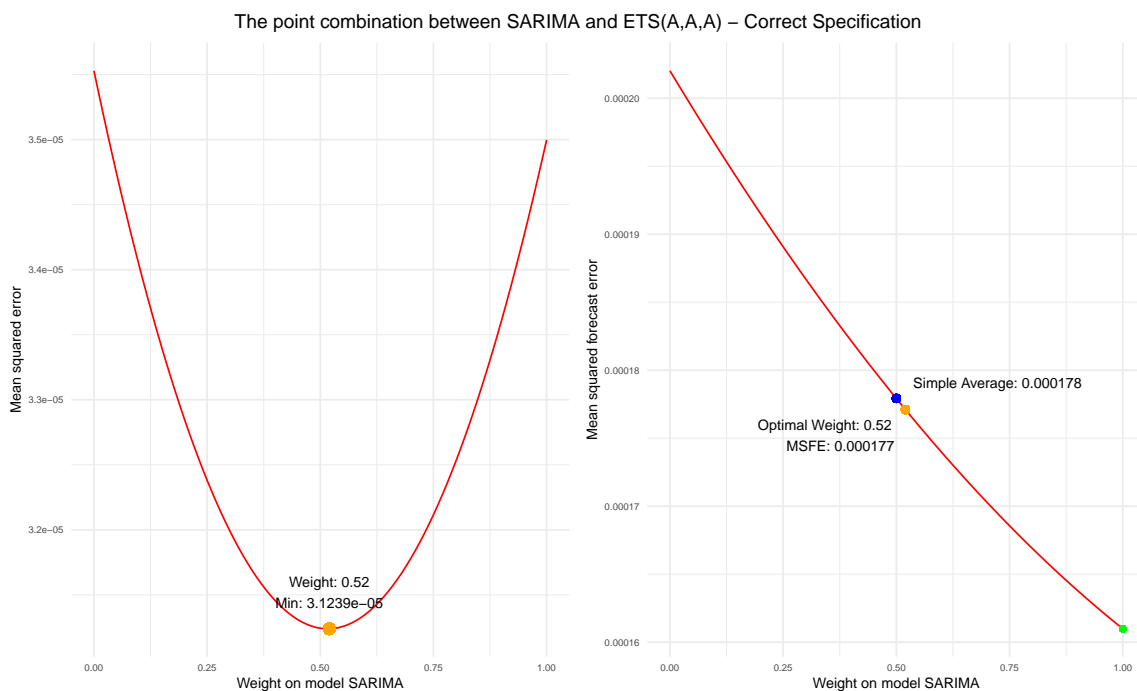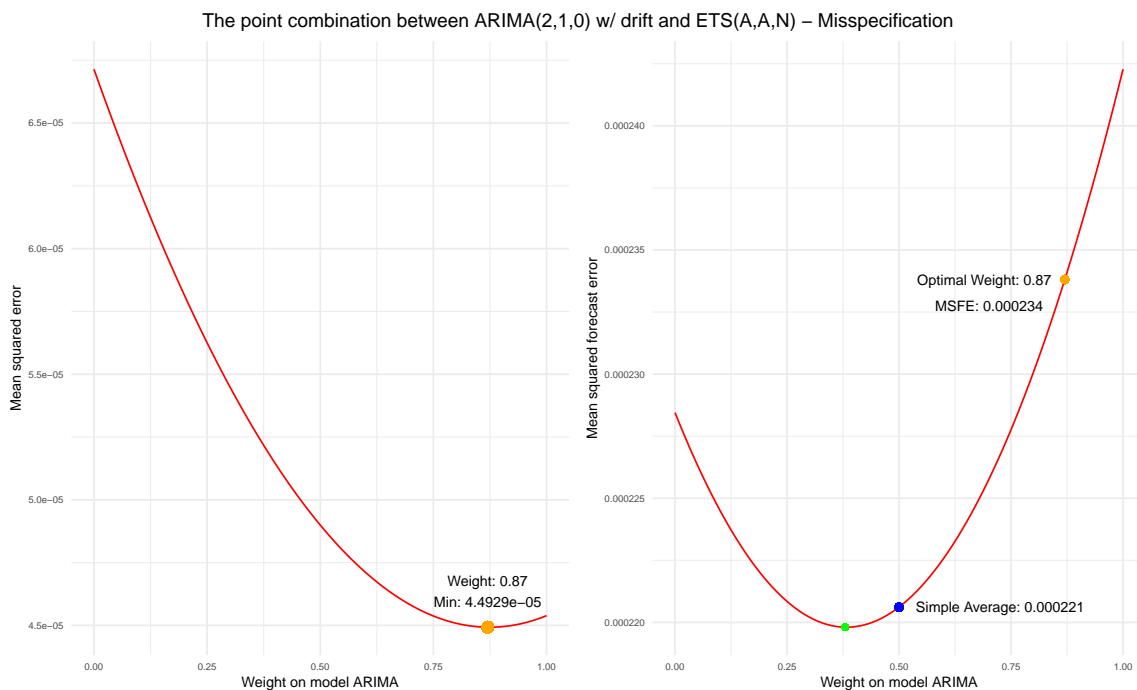

Figure 3.4: *MSFE of predictive points in correctly specified two-model pools over the in-sample (left) and out-of-sample (right) period. The x-axis represents the weight assigned on the SARMA model and the y-axis indicates the value of MSFE. The meanings of colored dots remain the same.*

The forecast combination puzzle is evident 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. From the in-sample combination accuracy plot, we can assert that both models fit the training set equally well, leading the estimated optimal weight to be around 0.5. These results exemplify that producing forecasts with correctly specified models in a two-model pool will lead to the forecast combination puzzle.

3.2.2 Misspecified models

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

Figure 3.5: *MSFE of predictive points in misspecified two-model pools over the in-sample (left) and out-of-sample (right) period. The x-axis represents the weight assigned on the SARMA model and the y-axis indicates the value of MSFE. The meanings of colored dots remain the same.*

Based on the left plot of Figure 3.5, the ETS model has a bigger mean squared error than the ARIMA model, indicating a poor model performance in fitting the training set. The difference between these misspecified models is much larger than those correctly specified models mentioned above. Furthermore, Figure 3.5 does not reveal any evidence of the puzzle, given the simple average performs superior than the optimal forecast combination. Although the

difference of MSFE, 0.000013, seems to be small in magnitude, comparing it with 0.000001 in the correct case, the difference is relatively conspicuous.

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. Rather than forecasting future value, cross-sectional data often helps to better understand the individual behavior and decision-making with changing attributes.

A simulated cross-sectional dataset is designed to study how related elements in the linear regression model affect the presence of the puzzle, as well as the performance of density combinations. In line with previous notations but under the cross-sectional setting, the subscript t will change to i to represent each individual observation.

CHECK MOTIVATION!!!!

4.1.1 Experimental design

The true data-generating process (DGP) is assumed to be a classic 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 15000 (N) artificial cross-sectional observations generated from 4.1 with $E[x_{1i}] = E[x_{2i}] = 0$, $Var(x_{1i}) = Var(x_{2i}) = 1$, $Cov(x_{1i}, x_{2i}) = 0.7$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (1, 2, 2)'$, $\mu_e = 5$ and $\sigma_e^2 = 10$.

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 misspecified models to generate density forecasts with each only contains one of the regressors. Assume Model 1 includes only x_{1i} as the regressor and the other model, Model 2, includes only x_{2i} as the regressor. The density forecast combinations will follow the construction of `two-model` pools and be evaluated by the log score functions.

- Sample size is $N = 15000$

- $E[x_{1i}] = E[x_{2i}] = 0$

- $Var(x_{1i}) = Var(x_{2i}) = 1$

- $Cov(x_{1i}, x_{2i}) = 0.7$

- The true value of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (1, 2, 2)'$

- $\mu_e = 5$ and $\sigma_e^2 = 10$

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 under 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 set of weights and the corresponding log predictive score in each case, while the blue dot indicates the forecast performance of the simple averaging method. The green dot, as a reference, refers to the maximum point of the out-of-sample curve.*

Figure 4.1 clearly reflects that when the sample size is large enough, the simple average of predicted densities, indicated by the blue dot, can retain the forecast accuracy with a small

difference in the log predictive score, compared with the optimal combination indicated by the orange dot. This is an evidence of facing forecast combination puzzle. Given the puzzle, we can change the true value of relevant elements one at a time while holding all others constant, and then summarize the conditions under which the puzzle is likely to be evident.

- **Sample Size**
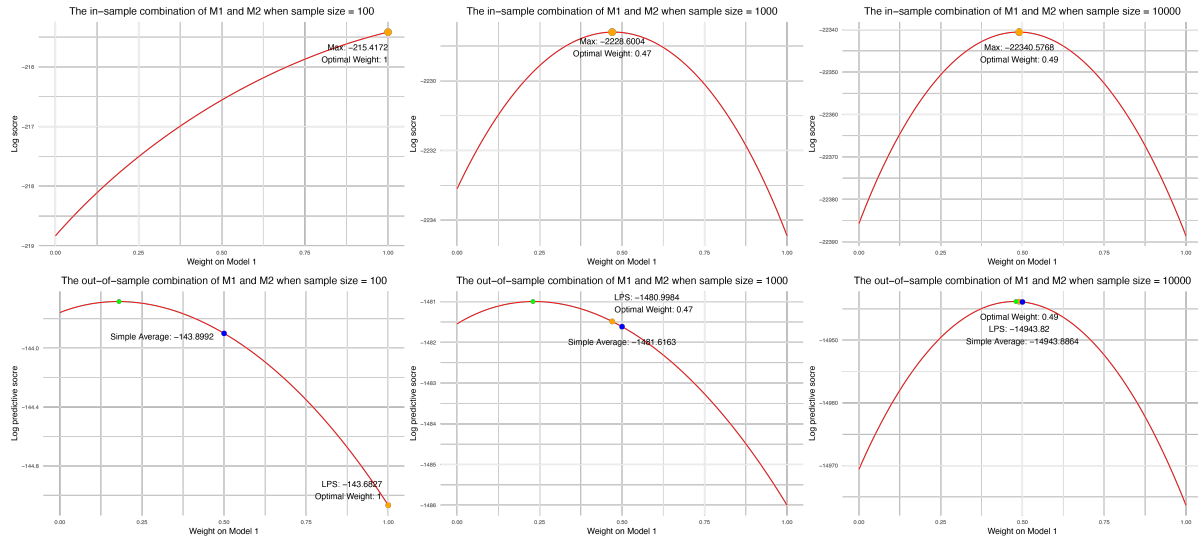
Figure 4.2: *Three columns refer to cases when $N = 100$, $N = 1000$, and $N = 10000$ respectively while keeping all others constant as the initial set-up. The top graphs represent the in-sample combination performance and the bottom graphs represent the out-of-sample combination accuracy. The meanings of colored dots are the same as those in Figure 4.1.*

First, it is notable that, in Figure 4.2, the performances of in-sample and out-of-sample combinations have completely different shapes or features when $N = 100$ but are gradually similar when $N = 1000$ and $N = 10000$. In the $N = 100$ case, we completely prefer Model 1 to fit the training set, however, the Model 1 becomes the worse choice for the test set. Thus, the averaged density forecast performs much better than the combination recommended by the optimal weight. This implies that the model combination which fits the in-sample well does not necessarily generate better forecasts when the sample size is small. Second, the forecast accuracy of optimal combination and that of simple averaging are getting closer when the sample size increases, which indicates the presence of the forecast combination puzzle. Roughly, the puzzle becomes noticeable when the sample size is larger than 500.

When we have a small dataset, it is not representative of the whole population, so the model estimation involves more randomness and is highly influenced by potential outliers. There is also a possibility of overfitting the training set when the training and test sets have distinct

patterns. Therefore, given a large enough dataset and two equally good models, we are very likely to have the puzzle.

- **Magnitude and Sign of β**

Next, the sample size is set to be 10000 so that it is large enough to reveal the puzzle. Consider the change in magnitude and sign of β_1 and β_2.
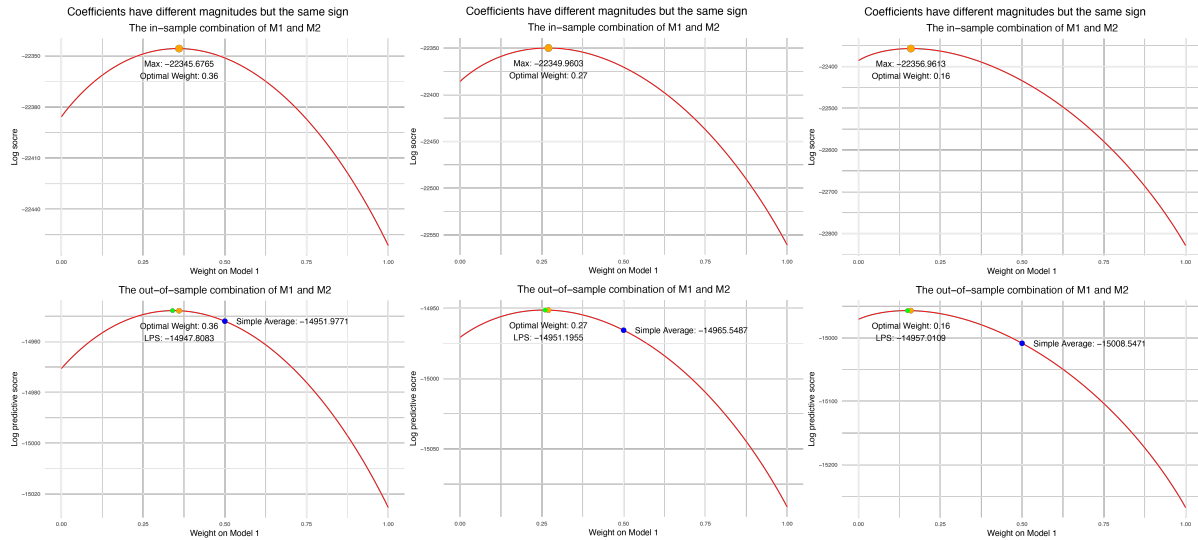

Figure 4.3: *In this case, β_1 and β_2 have the same sign but different magnitudes. The first column refers to $\beta_1 = 2$ and $\beta_2 = 3$, the second column refers to $\beta_1 = 2$ and $\beta_2 = 4$, and the first column refers to $\beta_1 = 2$ and $\beta_2 = 6$.*

Based on the results shown in Figure 4.3, 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 observe the puzzle and the optimal combination always wins with a higher log predictive score. The larger the absolute difference, the bigger the difference of two log predictive scores.

In the linear regression analysis, the magnitude of each coefficient represents the influence size of each regressor on the dependent variable. A large coefficient means that a change in the corresponding regressor affects 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 4.3. The effect of x_2i on y_i, β_2, is relatively larger than the effect of x_1i on y_i, β_1, so the Model 2 with x_2i only should be weighted higher in the combination.

Figure 4.3 illustrate that when β_1 and β_2 only have opposite signs, the puzzle seems to be insensitive. In both cases, the optimal combination and simple averaging forecast have very
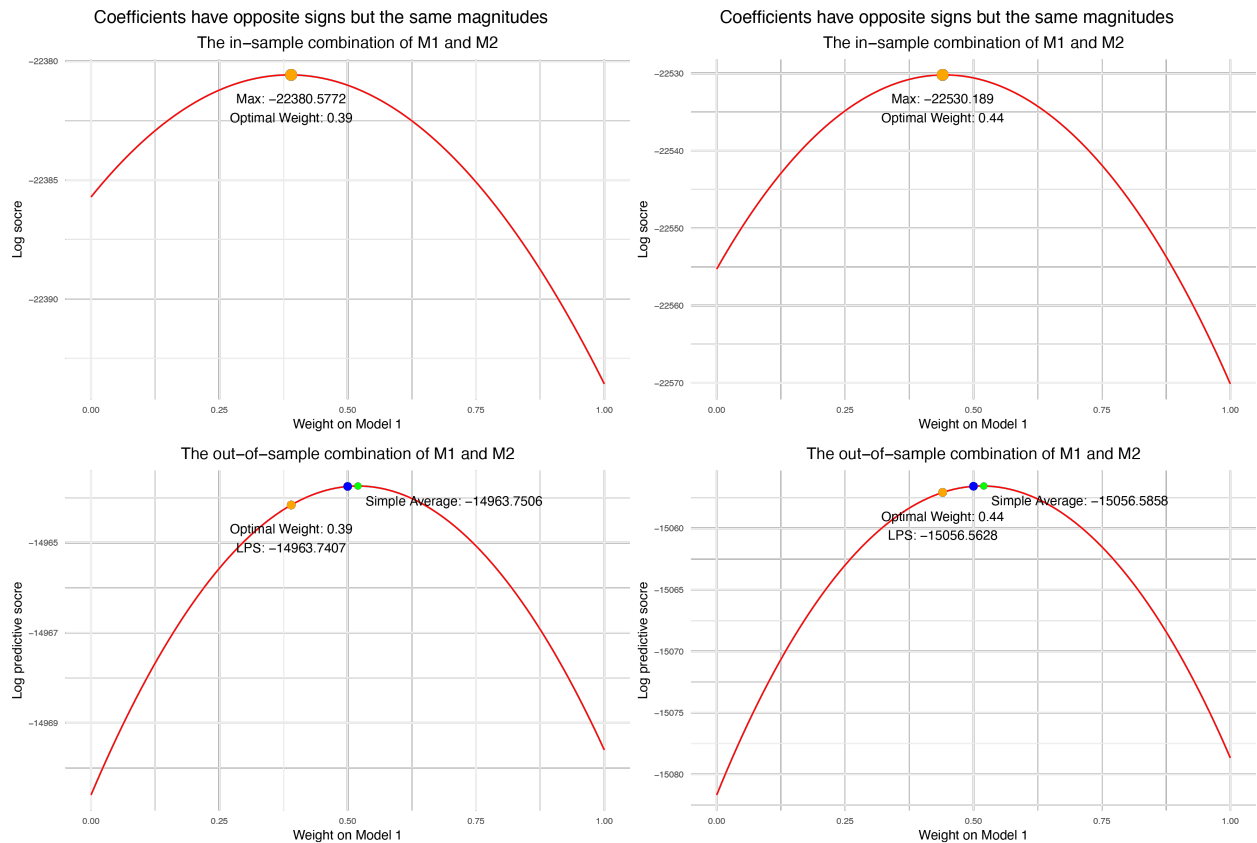
Figure 4.4: β_1 *and* β_2 *have the same magnitude but different signs, i.e.* $\beta_1 = -\beta_2$. *The first column considers the case when* $\beta_1 = 2$ *and* $\beta_2 = -2$ *and the second column considers the case when* $\beta_1 = 4$ *and* $\beta_2 = -4$.

similar log predictive scores, which is a strong evidence of the puzzle. Meanwhile, two regressors have the same effect in magnitude on y_i, therefore, the weight should be equally assigned in rough. We also notice that the accuracy of the optimal prediction combination can be improved by having higher absolute values of the coefficients. This makes regressors to have larger and more certain impacts on y_i, which is substantiated by the above case.

- **Variance of regressors**

larger variance, more variation to explain y, more favored

Variance of the error term will influence the log score value Pattern is hard to determine
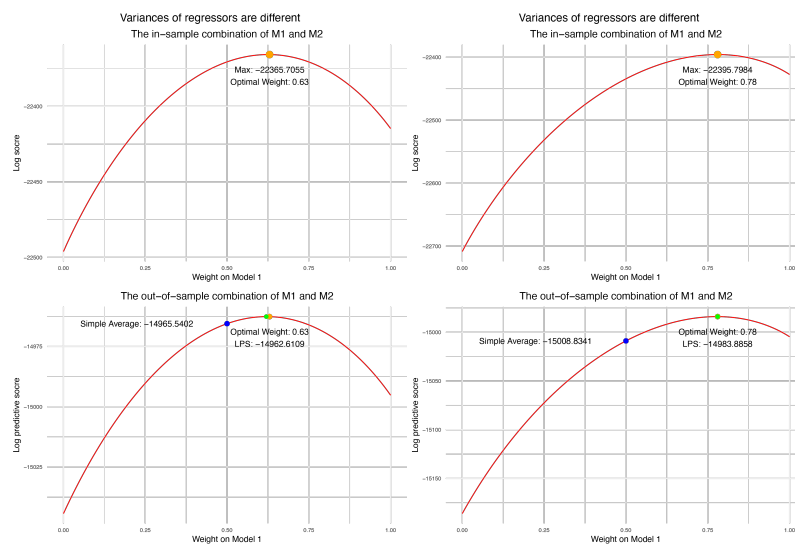
Figure 4.5: *The first column refers to the case when*

Discussion

5.1 Research Objective

Conclusion

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 forecasts is subject to future research.

This step, while important, is sizeable and we leave it to future work.

Appendix

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 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), and `fpp3` (Hyndman, 2023).

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).

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`.

Blanc, SM and T Setzer (2016). When to choose the simple average in forecast combination. *Journal of Business Research* **69**(10), 3951–3962.

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*.

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`.