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

October 2023

Contents

Abstract

This thesis demonstrates that the forecast combination puzzle is tightly related to the in-sample fit of constituent models used in the analysis. The forecast combination puzzle refers to the common finding that an equally-weighted forecast combination often outperforms an optimally-weighted forecast combination calculated via a sophisticated scheme. We show that when constituent models have similar in-sample fit, the puzzle will be in evidence; it is ambiguous otherwise. We analytically show the relationship between the estimated optimal weight and the constituent models in terms of point combinations using mean squared error and empirically confirm these findings in density combinations using log score. As an additional contribution, the puzzle is shown to be evident in both time series and cross-sectional settings.

Keywords: Forecast Combination, Forecast Combination Puzzle, Point Forecasting, Probabilistic Forecasting, Scoring Rules

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 those 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 relates to the in-sample performance of the constituent models used to produce the combination.

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 derive and obtain a closed-form expression that supports this finding in the linear 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 to forecasters and decision makers. The application of forecast combination, originally proposed in the seminal work of Bates and Granger (1969), provides improvements in forecast accuracy relative to individual forecasts, 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). Many researchers also devote efforts on probabilistic forecasting to obtain more information about the uncertainty of the resulting forecast. Similar to point forecasts, researchers have found that density forecast combination outperform 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 designed to improve in-sample accuracy should in theory perform better out-of-sample. However, the mean of the constituent 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. For the purposes of this analysis, we explicitly define the forecast combination puzzle as: 1) the simple average has better out-of-sample performance than that of the optimal combination; and 2) the forecast accuracy between the optimal combination and the simple average being small, which allows for the optimal combination to be slightly higher than the simple average. As the forecast accuracy is not too different, using either combination method will not make a meaningful difference, except in special circumstances.

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 overlook the possibility that complicated combination methods can perform better than the simple average in some cases. In order to make a rigorous explanation statement, we systematically explore the determinants behind the presence of the 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. Therefore, we will investigate the puzzle in both time series and cross-sectional settings using empirical and simulated data respectively.

Considering a simple case of two-model combination, our initial conjecture is that the presence of the puzzle is tightly-related to the in-sample fit of two constituent models. We conjecture that 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 greatly affects the behavior of forecasts, so forecasts produced by two similarly performed models will not differ much, leading to an estimated optimal weight around a half. Consequently, we should expect small differences in forecast accuracy between optimally-combined forecast and equally-combined forecast. It is then reasonable to prefer the simple average given that the forecast variance will also be lessened due to no extra parameter estimation. On the contrary, if two models have distinct in-sample fit, we conjecture that the optimal forecast combination will give more weight to the better performing forecast and therefore the estimated weights will be far away from a half. The estimated value of the optimal weight does not indicate the forecast accuracy of the optimal forecast combination. Therefore, there are two possible situations: the equally-combined forecast perform better than the optimally-combined forecast, and the opposite. The presence of the puzzle now becomes ambiguous, depending on the situation we fall into. According to the definition of the puzzle, our conjecture can be summarized in Table 1.1. Two constituent models are evaluated based on their in-sample relative performance and are also allowed 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 presence of the forecast combination puzzle, while "?" implies that the presence of the puzzle is uncertain.*

We demonstrate that the forecast combination puzzle is in evidence in the time series setting with the S&P500 index and the quarterly unemployment data. That is, the equally-weighted combination provides equivalent or superior forecast accuracy relative to the optimally-weighted combination. We then compare the in-sample fit of constituent models using their in-sample log likelihood and validate most of our conjecture.

Furthermore, we investigate the forecasting performance of two-model pools for simulated cross-sectional data using simple linear regression models. We derive a mathematical relationship between the optimal combination weight under the mean squared forecast error and elements in the true DGP. This relationship implies that the forecast combination puzzle is tightly-related to the interaction between constituent models and the true DGP. Given this knowledge in the point combination setting, we empirically investigate this finding for density combinations and validate that these formal reasoning are applicable in a more general setting. In addition, this simulation study provides sufficient empirical evidence to examine conjecture.

The goal of this thesis is two-fold: first, to substantiate the presence of the combination puzzle in the time series setting and to explore the relationship between the puzzle and the in-sample fit of constituent models; second, to mathematically derive the formula of the optimal weight in the regression setting under mean squared forecast error and then to validate the conjecture in Table 1.1 with empirical evidence.

The thesis follows two common weighting schemes for `two model` pools (Section 2) and then applies the log score to density combinations for daily S&P 500 index and the mean squared error to point combinations for quarterly number of unemployed (Section 3). With some empirical evidence of the conjecture, Section 4 derives a closed-form expression for the optimal weight under the mean squared error in a simple regression case. The findings are further examined by analyzing density combinations in the cross-sectional setting. The final section concludes.

Methodology

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;
2. combining forecasts based on a given accuracy criteria.

We only consider the combination of two individual forecasts, i.e., two constituent models, to simplify the analysis through fast and relative simple data manipulation.

Before explaining details, the following notation will be used throughout the paper. To examine forecast accuracy, we partition on observed time series y_t with a total of T observations into two proportional parts, an in-sample period with R observations and an out-of-sample period with P observations. We restrict the analysis to a 1-step ahead prediction, conditioning on the information set at time t, $\mathcal{F}_t$, which is comprised of all observed (and known) realizations of y up to time t, i.e., $\mathcal{F}_t = \{y_1, y_2, .., y_t\}$.

Every proposed parametric model determines the conditional probability density for y_t, denoted by $f(y_t|\mathcal{F}_{t-1}, \theta)$, given unknown parameters θ 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 maximum likelihood estimation method is applied to generate the estimators of unknown parameters, i.e., $\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{R} log f(y_t|\mathcal{F}_{t-1}, \theta)$. 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

Linear pooling

Consider the case of 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 weights 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:

$$f_\omega(y_t) = \omega\, f_1(y_t|\mathcal{F}_{t-1}, \theta_1) + (1-\omega) f_2(y_t|\mathcal{F}_{t-1}, \theta_2),$$

where $\omega \in [0,1]$ is the non-negative weight allocated to the probability density attributed to the first model. Two densities are often determined by different sets of parameters, differentiated by θ_1 and θ_2. Through this construction, the sum of the model weights is fixed at 1, which is a necessary and sufficient condition for $f_\omega(y_t)$ 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).

Log scoring rules

Following the literature on density evaluation, our initial analysis will focus on using 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, the log score over the in-sample period is:

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

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 R observations,

$$\hat{\omega}_{\text{opt}} = \underset{\omega\in[0,1]}{\arg\max} \sum_{t=1}^{R} log\Big[\omega\ \hat{f}_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_1) + (1-\omega)\ \hat{f}_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_2)\Big]. \tag{2.1}$$

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

$$LPS = \sum_{t=R+1}^{T} log\Big[\hat{\omega}_{\text{opt}}\, \hat{f}_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_1) + (1-\hat{\omega}_{\text{opt}})\, \hat{f}_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_2)\Big]. \tag{2.2}$$

2.2 Point combinations

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

Linear combination

Similar to the density case, points from two constituent models, y_{1t} and y_{2t}, are aggregated linearly:

$$y_t(\omega) = \omega\, y_{1t} + (1-\omega)\, y_{2t},$$

where $\omega \in [0,1]$ is the non-negative weight allocated to the point forecast generated from the first model.

Mean squared forecast error

The mean squared error (MSE) of the individual prediction 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:

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

The lower the MSE, the higher the accuracy of the forecast. Therefore, the "optimal" set of weights minimizes the MSE of the point forecast combination among all other possible sets over the R in-sample observations:

$$\hat{\omega}_{\text{opt}} = \underset{\omega\in[0,1]}{\arg\min}\frac{1}{R}\sum_{t=1}^{R}\Big[y_t - (\omega\, \hat{y}_{1t} + (1-\omega)\, \hat{y}_{2t})\Big]^2.$$

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

$$MSFE = \frac{1}{P} \sum_{t=R+1}^{T} \left[y_t - (\hat{\omega}_{\text{opt}}\, \hat{y}_{1t} + (1 - \hat{\omega}_{\text{opt}})\, \hat{y}_{2t}) \right]^2.$$

Empirical Results

3.1 Daily time series (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 $T = 2519$ total 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% ($P = 1008$) is reserved to evaluate forecast performance.

We will investigate the presence of the forecast combination puzzle when both models fit the training set well and when one model badly fits the data. Three predictive models are chosen to study the performance of density predictions across sets of two-model pools from common classes of linear time series models: autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and linear regression model with ARIMA errors (LR). Detailed model specifications are elaborated in the Appendix.

We use $P(A_1, A_2; \omega_{opt})$ to denote a two-model pool, where A_1 and A_2 are constituent models in the pool and ω_{opt} is the optimal weight assigned to the first model A_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 ω takes values on a grid from 0 to 1 with increment of 0.01. The log score, as a function of the weight ω, is generated to search for the optimal weight

over the in-sample period (refer to the top row of Figure 3.1). According to equation (2.1), 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 (2.2).

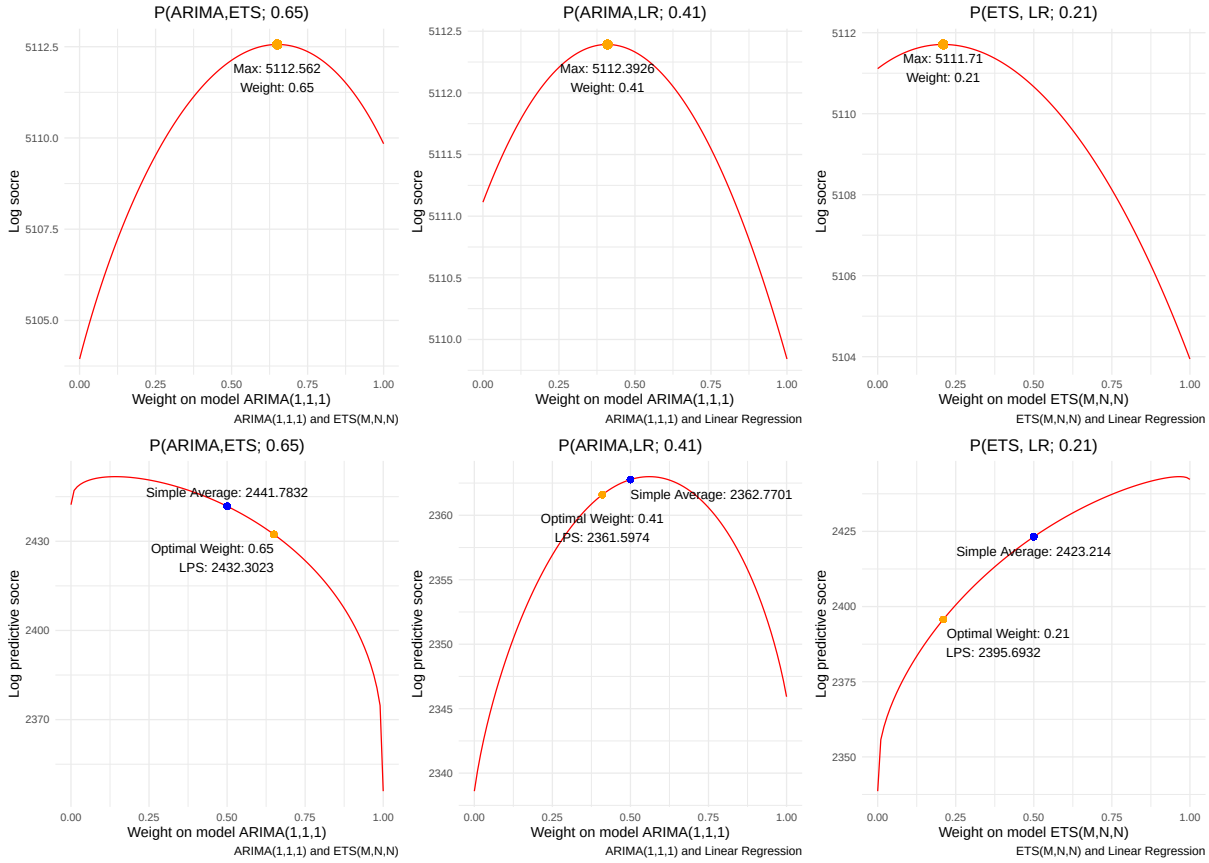

Figure 3.1: *Log predictive score of S&P 500 index predictive densities in two-model pools over the in-sample (top) and out-of-sample (bottom) period. Constituent prediction models are described in the title with 'P' representing 'Pool'. The x-axis represents the weight assigned on the former model of 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, i.e., the simple average performs better than the optimal combination. It is noticeable that the accuracy differences are different. For example, the accuracy difference in P(ARIMA,LR; 0.41) is much smaller than that in other two pools. This is tightly related to the in-sample fit of the constituent models, which can be represented by the log likelihood value. Table 3.1 illustrates the log likelihood values of constituent models in each pool and the absolute difference.

One explanation for the poor performance of the ETS(M,N,N) model could be that it fails to capture the trend component, as shown in Figure 3.2. Compared with ARIMA and linear

regression, the ETS model fits the training set poorly, making it a relatively `Bad` model in two-model pools as a consequence.

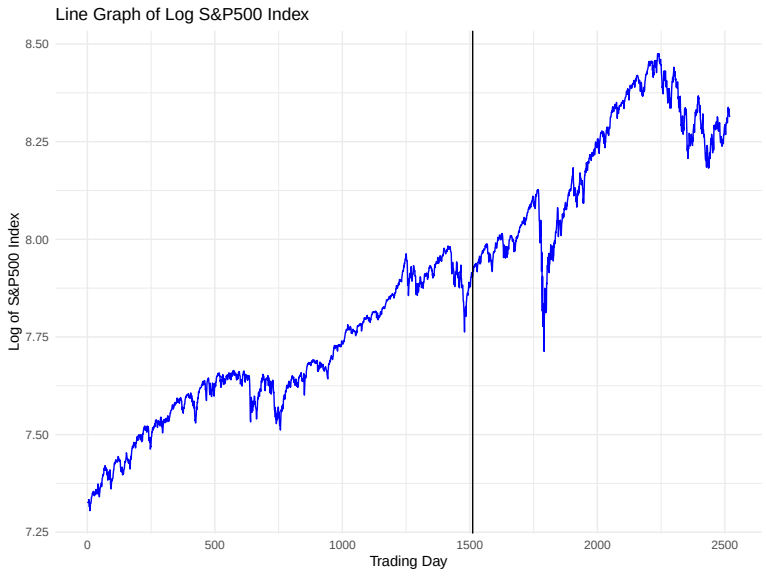

Figure 3.2: *The black vertical line separates the traning set and the evaluation set. The training set is on the left and the evaluation set is on the right.*

Linking it with our preliminary conjecture in Table 1.1, P(ARIMA,ETS; 0.65) can be viewed as a (B,G) case where the ARIMA model has a much better in-sample fit than the ETS mode. Similarly, we have a (G,B) case for P(ETS,LR; 0.21) since the LR model performs much better than the ETS model. P(ARIMA,LR; 0.41), on the other hand, exemplifies a (G,G) case since two models fit the in-sample data well.

	P(ARIMA,ETS; 0.65)	P(ARIMA,LR; 0.41)	P(ETS,LR; 0.21)
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
Type	(G,B)	(G,G)	(B,G)
Presence of the puzzle	Yes	Yes	Yes

Table 3.1: *"Log Likelihood Difference" represents the absolute difference of in-sample fit between two models, which is evaluated by the log likelihood. "Type" refers to the case of each two-model pool in the conjecture table. "Presence of the puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

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 series, which is covariance stationary. A series is said to

be covariance stationary when it has constant mean and variance, and the covariance between two observations at different time points depends on their time interval only.

Two candidate models are automatically selected by the `ARIMA()` function in the fable package (Hyndman, 2023): a Gaussian ARMA(1,1) model and a linear regression model with ARMA(1,1) errors. To differentiate with the first LR model, the second model is named as Linear Regression 2 (LR2) in the pool. Figure 3.3 illustrates that two constituent models have a very similar in-sample log score with only 0.0011 difference, and the puzzle is evidenced by the only 0.1282 accuracy difference between two forecast combination approaches. Strictly speaking, the forecast accuracy of the optimally-weighted combination (2349.764) is slightly superior to that of equally-weighted combination (2349.636). However, this difference in forecast accuracy is so small as to make no real difference in practice. In addition, using equal weights is much more efficient than estimating optimal weights through any weighting scheme. Recall that our definition of the forecast combination puzzle are cases where 1) the simple average performs much better out-of-sample than the optimal combination; and 2) the forecast accuracy of the two methods is similar. Comparing the log likelihood of two models in Table 3.2, the similar in-sample performance is another evidence of having a (G,G) case with the puzzle in evidence.

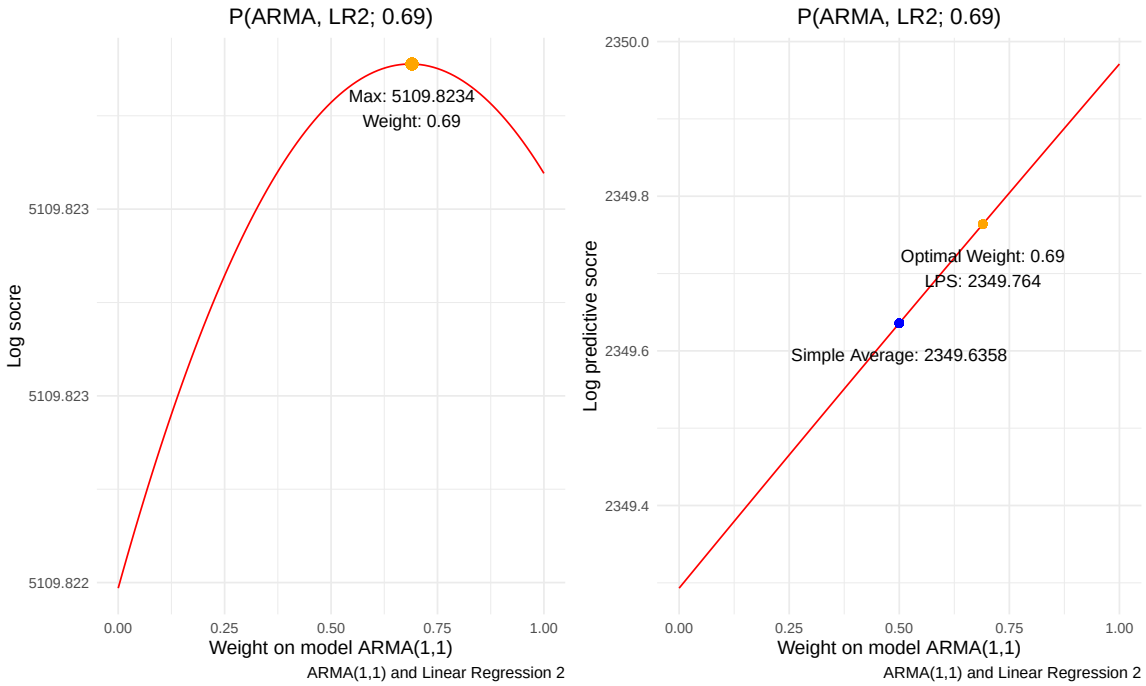

Figure 3.3: *Log predictive score of S&P 500 log returns predictive densities in two-model pools over the in-sample (left) and the out-of-sample (right) period. The x-axis represents the weight assigned on the ARMA(1,1) model and the y-axis indicates the log predictive score. The meanings of colored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

	P(ARMA,LR2;0.69)
First Model Log Likelihood	5109.8071
Second Model Log Likelihood	5109.8054
Log Likelihood Difference	0.0016
Type	(G,G)
Presence of the puzzle	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. "Type" refers to the case of two models in the conjecture table. "Presence of the puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

This Section 3.1 provides some empirical evidence for our initial conjecture. When both models fit the data well, i.e., they are `Good` models, then the accuracy of the optimal density forecast combination is close to that of the average density forecast, 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 Seasonal time series

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 in-sample fit) or poorly-specified (poor in-sample fit). 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). We use the natural logarithm of the total number of unemployment to reduce the level of variability in the series.

It has $T = 153$ total observations and is slit into two sets in proportion. As before, the first 60% of the data ($R = 91$), as the in-sample period, is used to estimate all unknown parameters. The remaining 40% ($P = 62$) is the out-of-sample period reserved for forecast performance evaluation.

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 the ARIMA model. 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 in Figure 3.4; the accuracy difference between two combinations is negligible. The optimally-weighted point combination has a MSFE of 0.000177 and the equally-weighted forecast has a MSFE of 0.000178. Looking at the in-sample combination plot, two models fit the training set equally well, which can also be confirmed by the second column of Table 3.3.

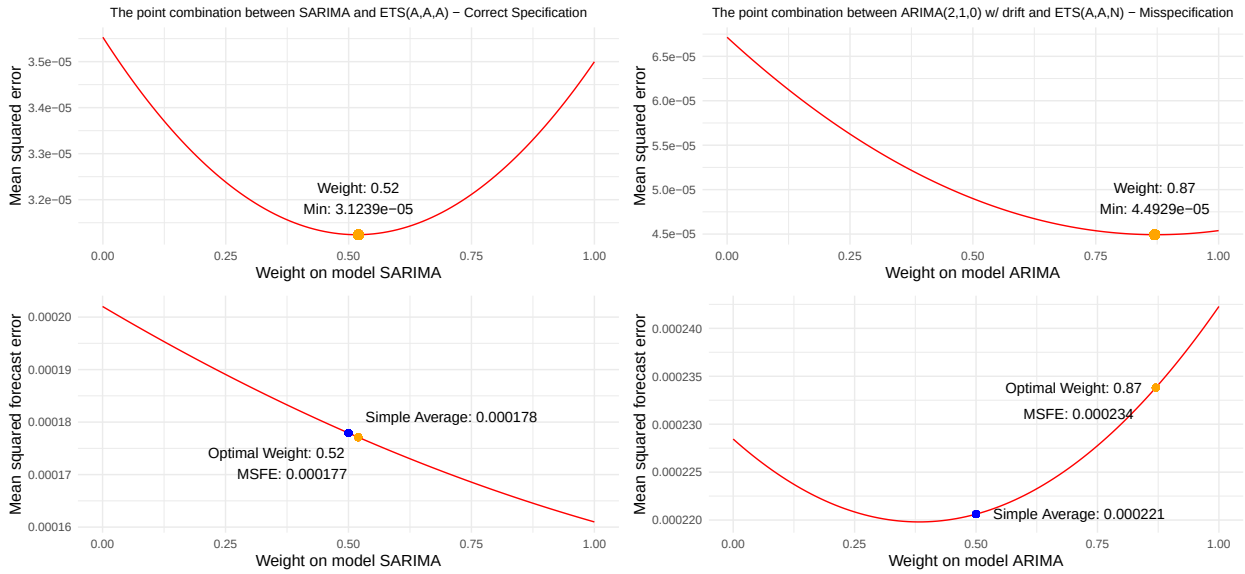

Figure 3.4: *MSFE of predictive number of unemploymed people in well-specified (left) and pooly-specified (right) two-model pools over the in-sample (top) and out-of-sample (bottom) period. The x-axis represents the weight assigned on the first model and the y-axis indicates the value of MSFE. The meanings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

Poorly-specified models

One way of proposing a `Bad` model for a seasonal dataset is to deliberately ignore the seasonality in data. In this section, we consider the ARIMA(2,1,0) and ETS(A,A,N) models, which do not capture any seasonal component and are the best models automatically selected by `ARIMA` and `ETS` functions in R.

	P(SARIMA,ETS;0.52)	P(ARIMA,ETS;0.87)
First Model Log Likelihood	321.4497	322.1642
Second Model Log Likelihood	260.9102	231.9507
Log Likelihood Difference	60.5395	90.2135
Type	(G,G)	(B,B)
Presence of the puzzle	Yes	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. "Type" refers to the case of two models in the conjecture table. "Presence of the puzzle" indicates whether the simple average is close to or outperforms the optimal forecast combination.*

The right column of Figure 3.4 does reveal the forecast combination puzzle, as the equally-weighted combination performs better than the optimally-weighted forecast combination. Furthermore, the third column of Table 3.3 illustrates that both models have similar log likelihood. We may claim that, regardless of whether the constituent models capture all the features of the data, as long as they have similar in-sample performance, the forecast combination puzzle will be evidenced. As a result, we should also expect to find the puzzle if two models are equally `Bad`.

The empirical evidence suggest that the puzzle is in evidence in all cases, however, these examples are not enough to draw comprehensive conclusions. Also, one big challenge of working with empirical data is that the true DGP is unknown.

To partially circumvent these issues, we use a simulation experiment based on a pure cross-sectional process to further investigate when the optimal forecast combination is more accurate than the simple averaging. Compared with simulated time series data, there is no need to consider the dependence of observations, hence, is an easy starting point. As an additional contribution, it examines the presence of the forecast combination puzzle in the cross-sectional setting.

Pure Cross-sectional Analysis

Given that forecast combinations can greatly improve forecast accuracy, this idea can also be applied to the pure cross-sectional settings. Remark that the use of prediction combinations is not common in such setting. In this section, we derive an analytical closed-form expression of the optimal weight under mean squared forecast error to investigate the determinants behind the puzzle in the cross-sectional setting. A simulation study is then conducted to evaluate and verify the applicability of findings.

Compared with real-life data, implementing simulation is easy to control and interpret given that the true DGP is known. Meanwhile, it is an effective way of validating our conjecture by freely changing the elements of true DGP and looking for the forecast combination puzzle. 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 Model Setup

The true DGP is assumed to be a linear regression model with no intercept and only two exogenous and weakly correlated regressors, which satisfy all assumptions of the classical linear regression model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \;\; \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2) \quad (i = 1, 2, ..., N). \tag{4.1}$$

The forecasting models, or constituent models, are

$$M_1 : y_i = \alpha_1 x_{1i} + e_{1i}, \;\; e_{1i} \overset{i.i.d}{\sim} N(0, \sigma_1^2)$$
$$M_2 : y_i = \alpha_2 x_{2i} + e_{2i}, \;\; e_{2i} \overset{i.i.d}{\sim} N(0, \sigma_2^2).$$

4.2 Optimal Weight (MSE)

Following the methodology in Section 2, the observed data are divided into an in-sample period (R) for parameter estimation and an out-of-sample period (P) for accuracy evaluation. As noted before, the estimated optimal weight, $\hat{\omega}_{opt}$, will be generated using the first R number of observations, and held fixed over the P observation.

To simplify the notation, we use the notation for linear regression models in matrix form, and obtain the following formula for $\hat{\omega}_{opt}$ under the MSE loss

$$\hat{\omega}_{opt} = \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'x_1'x_1\hat{\alpha}_1 - 2\hat{\alpha}_1'x_1'x_2\hat{\alpha}_2 + \hat{\alpha}_2'x_2'x_2\hat{\alpha}_2},$$

where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the ordinary least squares estimators in M_1 and M_2 respectively.

A more meaningful expression can be achieved by multiplying $\frac{1}{R}$ to both numerator and denominator and writing

$$\hat{\omega}_{opt} = \frac{\hat{\alpha}_1'\text{cov}_R(x_1, x_1)\hat{\alpha}_1 - \hat{\alpha}_1'\text{cov}_R(x_1, x_2)\hat{\alpha}_2}{\hat{\alpha}_1'\text{cov}_R(x_1, x_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\text{cov}_R(x_1, x_2)\hat{\alpha}_2 + \hat{\alpha}_2'\text{cov}_R(x_2, x_2)\hat{\alpha}_2},$$

where $\text{cov}_R(x_j, x_k)$ is the in-sample covariance between regressors x_j and x_k.

In the classical linear regression setting, OLS estimator is consistent when the sample size goes to infinity. That is, we should have $\hat{\alpha}_1 \xrightarrow{p} \alpha_1$ and $\hat{\alpha}_2 \xrightarrow{p} \alpha_2$. Considering the limit result, we have

$$\hat{\omega}_{opt} \xrightarrow{p} \omega_\star = \frac{\alpha_1'\Sigma_{11}\alpha_1 - \alpha_1'\Sigma_{12}\alpha_2}{\alpha_1'\Sigma_{11}\alpha_1 - 2\alpha_1'\Sigma_{12}\alpha_2 + \alpha_2'\Sigma_{22}\alpha_2}, \tag{4.2}$$

where $\omega_\star$ is the limiting value of the optimal weight, Σ_{jk} denotes the population covariance matrix of corresponding regressors x_j and x_k. With the limit result in equation (4.2), we can easily work out the asymptotic determinants of having $\omega_\star = \frac{1}{2}$ and then connect it with the presence of the puzzle.

For $\omega_\star = \frac{1}{2}$, it must be that $\alpha_1'\Sigma_{11}\alpha_1 = \alpha_2'\Sigma_{22}\alpha_2$, which suggests a symmetrical relationship between two constituent model. This gives rigorous evidence that similar in-sample performance of two models will lead to the presence of the puzzle. Besides, any situation where this final equality is nearly satisfied will inevitably lead an optimal weight near one-half.

In addition, according to the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in Appendix A.2, α_1 and α_2 will be close to β_1 and β_2, respectively, when the correlation between regressors is small. This suggests that the optimal weight interacts with the true data generating process and is therefore related to the true DGP.

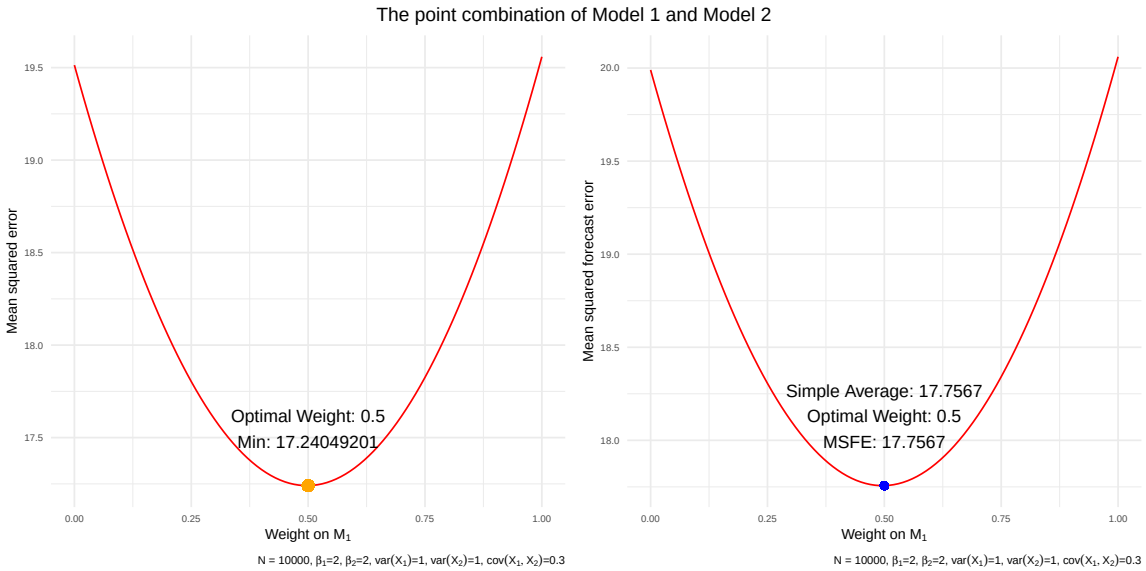

Figure 4.1: *MSFE of predictive points in a two-model pool over the in-sample (left) and the out-of-sample (right) period. The x-axis represents the weight assigned on M_1 and the y-axis indicates MSFE. The orange dot represents the optimally-weighted combination, while the blue dot indicates the equally-weighted combination.*

Figure 4.1 illustrates one example of having $\hat{\omega}_{opt}$ equal to 0.5 when $N = 10000$, $\beta_1 = \beta_2 = 2$, $Var(X_1) = Var(X_2) = 1$, $Cov(X_1, X_2) = 0.3$ under the MSE weighting scheme. Both in-sample and out-of-sample curves look symmetric.

4.3 Density Simulations

Recall the equation (2.1) in Section 2, it is clear that the weight ω appears in the natural logarithm function. The expectation operator can only evaluate the first derivative of equation (2.1) with respect to ω by integration. As a consequence, there is no closed-form limiting expression for the optimal weight $\hat{\omega}_{opt}$ under log score. However, we can use a simulation study to examine the applicability of findings from Section 4.2 to density combinations and log scoring rules.

The initial set-up has 10000 (N) artificial cross-sectional observations generated from the equation (4.1) with $E[X_{1i}] = E[X_{2i}] = 0$, $Var(X_{1i}) = Var(X_{2i}) = 1$, $Cov(X_{1i}, X_{2i}) = 0.3$, $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (2, 2)'$, and $\sigma_\epsilon^2 = 4$. Same as before, around 60% of the data will be used for model estimation.

The density forecast combinations will follow the construction of two-model pools and be evaluated using the log score.

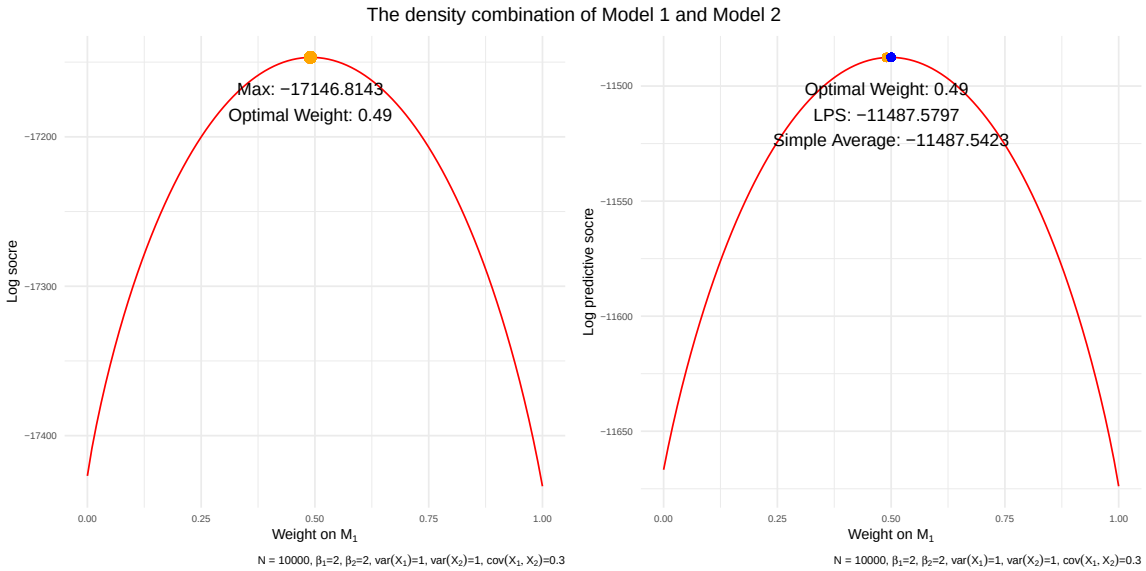

Figure 4.2: *Two curves refer to the in-sample (left) and out-of-sample (right) performance of density combinations with artificial cross-sectional data based on the initial set-up. The x-axis represents the weight assigned on M_1 and the y-axis indicates the log score. The meanings of colored dots remain unchanged.*

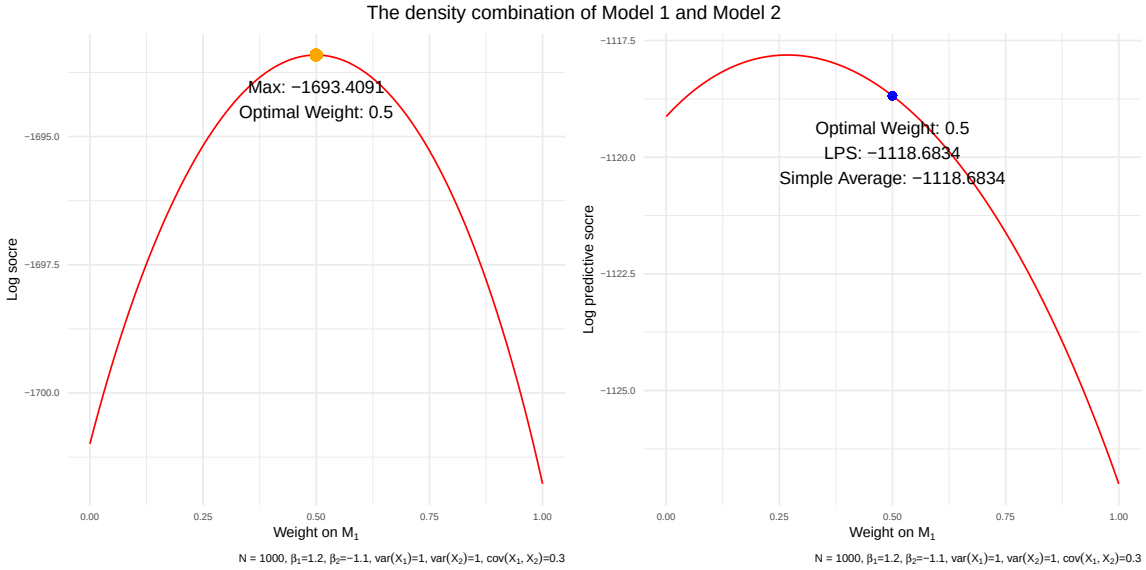

Figure 4.3: *Two curves refer to the in-sample (left) and out-of-sample (right) performance of density combinations with artificial cross-sectional data. The x-axis represents the weight assigned on M_1 and the y-axis indicates the log score. The meanings of dots remain unchanged.*

Using the same model specification as in Figure 4.1 to construct the density forecast combination via the log score, we see that the optimal weight $\hat{\omega}_{opt}$ is no longer one-half in Figure 4.2. On the other hand, Figure 4.3 shows that $\hat{\omega}_{opt}$ is equal to one-half when $N = 1000$, $\beta_1 = 1.2$, $\beta_2 = -1.1$.

This indicates that besides the magnitude of $\boldsymbol{\beta}$, the sign will also have an impact on $\hat{\omega}_{opt}$, even in large samples ($N = 50000$). Therefore, we speculate that the optimal weight has a non-linear relationship with the proposed models under the log score weighting scheme.

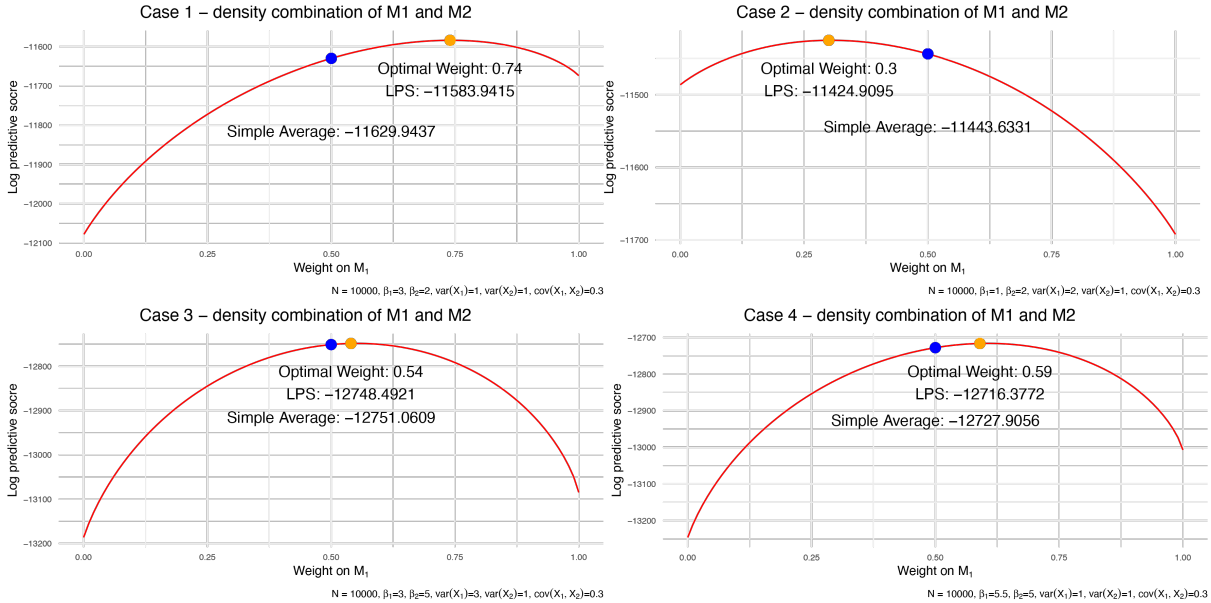

Figure 4.4: *The out-of-sample performance of density combinations with artificial cross-sectional data. The true DGP are in the bottom-right of individual plots. "Optimal Weight" shows the estimated optimal weight. The log predictive scores of optimally-weighted combination and equally-weighted combination are indicated by "LPS" and "Simple Average" respectively.*

In the regression setting, we evaluate the in-sample performance with R^2, and then link this value to the conjecture in Table 1.1. Four cases in Figure 4.4 are selected to further validate different types of model combinations as empirical evidence. Clearly, the optimally-weighted combination outperforms the equally-weighted combination with higher forecast accuracy in the first two cases; hence, we do not have the forecast combination puzzle. After comparing the in-sample fit of constituent models with R^2, these two pools can be labelled as (G,B) and (B,G), respectively, as summarized in Table 4.1. This supports our preliminary conjecture that when two models have differing levels of in-sample performance, the presence of the puzzle is unclear, i.e., the puzzle may or may not occur.

Recall that in our definition of the forecast combination puzzle, small accuracy differences make it hard to decide whether the puzzle is in evidence. For example, cases 3 and 4 in Figure 4.4 both illustrate a close distance between the optimal forecast combination and the simple average. Even the R^2 of constituent models in each pool are similar, as indicated by Table 4.1. One possible solution is to formally test the statistical significance of the accuracy difference through

	Case 1	Case 2	Case 3	Case 4
R^2 of M_1	0.393	0.141	0.476	0.558
R^2 of M_2	0.256	0.224	0.452	0.504
R^2 Difference	0.138	0.083	0.024	0.053
Type	(G,B)	(B,G)	(B,B)	(G,B)
Presence of the puzzle	No	No	Yes	No

Table 4.1: *"Difference" is the absolute difference of in-sample R^2 between two models. "Type" refers to the case of two models in the conjecture table. "Presence of the puzzle" indicates whether the equally-weighted combination is close to or outperforms the optimally-weighted combination.*

the Diebold-Mariano Test (Diebold, 2015). Unfortunately, it has been proven that the test is not appropriate in this context as the test statistic will not follow a standard normal distribution using the two-stage estimation (Frazier et al., 2023). Another possible choice is deciding on an arbitrary value for the accuracy difference. It turns out that the magnitude of the log predictive score is tightly related to the sample size and assumptions about the error term in the true DGP. Hence, this method can be used when fixing the sample size and assumptions of the error term.

Instead of forecast accuracy, we can look at the difference in in-sample R^2. One possible rule of thumb is that when the absolute difference of in-sample R^2 between two models is less than 0.05, i.e., two models have similar in-sample fit, then the two-model pool can be viewed as either the (G,G) or (B,B) case, and therefore the puzzle seems to be in evidence. Applying this finding to cases 3 and 4 in Table 4.1, case 3 is a (G,G) case where its in-sample R^2 difference is 0.024, less than 0.05, whereas case 4 is a (G,B) case with a 0.053 in-sample R^2 difference, slightly higher than 0.05. By using this rule of thumb, we can only be certain that the puzzle is evident in case 3 but we are unconfident in case 4. In terms of the log likelihood, it is highly affected by the sample size, similar to the log predictive score. One possible way, however, is to normalize the difference of two log likelihoods by their sum based on the chosen constituent models. The heuristic is around 0.009, meaning that when the normalized difference is less than 0.009, two models have similar in-sample fit, and therefore the puzzle is likely to be in evidence.

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

Table 4.2: *The first row and the first column refer to two constituent models, M_1 and M_2. "Good" and "Bad" denote the relative in-sample fit of constituent model. The "$\surd$" indicates the presence of the forecast combination puzzle, while "?" implies that the presence of the puzzle is ambiguous.*

The analysis in this section provides a general idea of the relationship between the in-sample fit of constituent models and the presence of the puzzle in both point and density forecast combinations. Based on new empirical evidence, the conjecture table for a two-model pool remains the same but with updated definitions, as illustrated in Table 4.2.

Conclusion

This thesis develops a means of determining the presence of the forecast combination puzzle in a two-model pool by closely examing in-sample model fit. Empirical results suggest that when both constituent models have similar in-sample fit, poor or good, the equally-weighted combination will provide equivalent forecast accuracy relative to the optimally-weighted combination. On the other hand, when the in-sample fit differs between the two models, the presence of the puzzle is ambiguous.

Importantly, in a linear regression context, we derive the relationship between the optimal weight and elements in the proposed models under mean squared error scheme when using point combinations. According to the closed-form expression for the estimated optimal weight, the presence of the puzzle is tightly-related to the sample size, the sign and magnitude of parameters in the constituent models, the sample variances of regressors, and the correlation between regressors. It is also shown that a large difference in the in-sample performance of proposed models can move the estimated optimal weight away from one-half (equal weights), especially in large samples. Additionally, the optimal weight also interacts with the true DGP in a broad sense, which determines the estimated coefficients in the constituent models.

Not surprisingly, these findings can be applied to density combinations under log predictive score, where the estimated optimal weight does not have a closed-form expression. While the two constituent models have a symmetrical relationship when the population weight is equal to one-half in the MSE case, we find empirical evidence that this may not be the same case under log score.

Working with the two-model pools provides an opportunity to explore a variety of situations in a short period of time. The next natural step is to investigate multiple forecast combinations. It is also necessary to relax some of the restrict model assumptions and increase the complexity

of the model structure in our simulation study. Under the explicit definition of the forecast combination puzzle, it is hard to determine the significance of the accuracy difference between optimally-weighted combination and equally-weighted combination, given that neither the testing or an arbitrary choice will work for all cases. We leave these, and other interesting issues, for 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, ϵ_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. The index t takes the values from 1 to the total sample size T.

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.

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

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

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

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.

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

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.

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

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

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

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 Optimal Weight Derivation Details

In this section, we detail the derivation steps of producing the results in Section 4.2.

Recall that the data is drawn from the true DGP:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \;\; \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2) \quad (i = 1, 2, ..., N),$$

where the in-sample period (R) is used to estimate the parameters for the following two constituent models:

$$\begin{aligned} M_1 : y_i &= \alpha_1 x_{1i} + u_{1i}, \;\; u_{1i} \overset{i.i.d}{\sim} N(0, \sigma_1^2) \\ M_2 : y_i &= \alpha_2 x_{2i} + u_{2i}, \;\; u_{2i} \overset{i.i.d}{\sim} N(0, \sigma_2^2). \end{aligned}$$

Besides, we allow x_{1i} and x_{2i} to have a small correlation, otherwise, there may be multicollinearity, resulting in higher standard errors of estimated parameters.

For simplicity, these models can be written in matrix form

$$\begin{aligned} y &= x_1\beta_1 + x_2\beta_2 + \epsilon, \\ M_1 : y &= x_1\alpha_1 + u_1, \\ M_2 : y &= x_2\alpha_2 + u_2, \end{aligned}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \; x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix}, \; x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{N2} \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

Applying the OLS estimation, we can immediately obtain the formula of $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Given a weak correlation between regressors, each formula will have an extra component, which represents that correlation.

$$\begin{aligned} \hat{\alpha}_1 &= (x_1'x_1)^{-1}x_1'y \\ &= (x_1'x_1)^{-1}x_1'(x_1\beta_1 + x_2\beta_2 + \epsilon) \\ &= \beta_1 + (x_1'x_1)^{-1}x_1'x_2\beta_2 \\ &= \beta_1 + var(x_1)^{-1}cov(x_1, x_2)\beta_2 \end{aligned}$$

$$\begin{aligned} \hat{\alpha}_2 &= (x_2'x_2)^{-1}x_2'y \\ &= (x_2'x_2)^{-1}x_2'(x_1\beta_1 + x_2\beta_2 + \epsilon) \\ &= \beta_2 + (x_2'x_2)^{-1}x_2'x_1\beta_1 \\ &= \beta_2 + var(x_2)^{-1}cov(x_2, x_1)\beta_1 \end{aligned}$$

$$\begin{aligned} \hat{y}_\omega &= \hat{y}_1\omega + \hat{y}_2(1-\omega) \\ &= x_1\hat{\alpha}_1\omega + \; x_2\hat{\alpha}_2(1-\omega) \\ &= x_1\hat{\alpha}_1\omega - x_2\hat{\alpha}_2\omega + x_2\hat{\alpha}_2 \\ &= (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega + x_2\hat{\alpha}_2 \end{aligned}$$

$$\begin{aligned} \hat{\omega}_{\text{opt}} &= \underset{\omega\in[0,1]}{\arg\min} \frac{1}{R}(y - \hat{y}_\omega)'(y - \hat{y}_\omega) \\ &= \underset{\omega\in[0,1]}{\arg\min} \frac{1}{R}[y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega - x_2\hat{\alpha}_2]'[y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega - x_2\hat{\alpha}_2] \end{aligned}$$

$$
\begin{aligned}
-2(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\hat{\omega}_{opt} - x_2\hat{\alpha}_2) &= 0 \\
(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\hat{\omega}_{opt} &= (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2) \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2)}{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)} \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2)}{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)} \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'x_1'x_1\hat{\alpha}_1 - 2\hat{\alpha}_1'x_1'x_2\hat{\alpha}_2 + \hat{\alpha}_2'x_2'x_2\hat{\alpha}_2}
\end{aligned}
$$

A.2.1 Formula related to the F-statistics

To clearly see the relationship between the in-sample fit and the optimal weight, the equation (4.2) in Section 4.2 can be linked with the F-statistics of two models. The F-test of overall significance is a formal hypothesis test, which examines the explanatory power of the whole model.

The hypothesis of the overall significance test for M_1 can be written as $H_0 : R\alpha_1 = r$ and $H_1 : R\alpha_1 \neq r$ where R is a scalar 1 (or an identity matrix when α is a column vector) and r is a scalar 0 (or a column vector of 0).

Define m as the number of restrictions in the null hypothesis, and the sum squared of errors (SSE) for the full (true) model 4.1 is $SSE_{full} = (y - x_1\hat{\beta}_1 - x_2\hat{\beta}_2)'(y - x_1\hat{\beta}_1 - x_2\hat{\beta}_2)$. Then the unbiased estimator of the true model variance is $s^2 = \frac{SSE_{full}}{R-2}$.

The F-statistic follows a F-distribution with degrees of freedom (1,R-2) under H_0, which is defined as

$$
\begin{aligned}
F_{\alpha_1} &= (R\hat{\alpha}_1 - r)'[s^2R(x_1'x_1)^{-1}R']^{-1}(R\hat{\alpha}_1 - r)/m \\
&= (\hat{\alpha}_1 - 0)'\left[s^2(x_1'x_1)^{-1}\right]^{-1}(\hat{\alpha}_1 - 0)/1 \\
&= R\, s^{-2}\, \hat{\alpha}_1'\mathrm{cov}_R(x_1, x_1)\hat{\alpha}_1.
\end{aligned}
$$

Similarly, we have

$$
F_{\alpha_2} = R\, s^{-2}\, \hat{\alpha}_2'\mathrm{cov}_R(x_2, x_2)\hat{\alpha}_2 \sim F_{1,R-2} \text{ under } \mathrm{H}_0.
$$

The optimal weight can also be constructed by the F-statistics of M_1 and M_2.

$$\hat{\omega}_{opt} = \frac{F_{\alpha_1} - R\ \hat{\alpha}_1' \text{cov}_R(x_1, x_2)\hat{\alpha}_2/s^2}{F_{\alpha_1} + F_{\alpha_2} - 2R\ \hat{\alpha}_1' \text{cov}_R(x_1, x_2)\hat{\alpha}_2/s^2}.$$

If the covariance between x_1 and x_2 is close to zero, the optimal weight can be approximated as $\hat{\omega}_{opt} = \frac{F_{\alpha_1}}{F_{\alpha_1}+F_{\alpha_2}}$. This is the reason why the in-sample performance of model is highly correlated with the presence of the puzzle.

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.

Diebold, FX (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics* **33**(1), 1–1.

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

Frazier, DT, R Covey, GM Martin, and D Poskitt (2023). Solving the Forecast Combination Puzzle. *arXiv preprint arXiv:2308.05263*.

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.

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.