BERGISCHE
UNIVERSITÄT
WUPPERTAL

Schumpeter School
of Business and Economics

Forecast Combination with Constrained Weights

Inauguraldissertation zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaft (doctor rerum oeconomicarum)
an der Fakultät für Wirtschaftswissenschaft
– Schumpeter School of Business and Economics –
der Bergischen Universität Wuppertal

vorgelegt von
Lars Averkamp, geb. Wißmann, M. Sc.
aus Wuppertal

Wuppertal, 01. Juni 2024

For my wife and unborn daughter

Acknowledgment

I would like to express my heartfelt gratitude to my wife Jutta Averkamp for her unwavering support throughout this journey. I am deeply thankful to my dissertation buddy and sister-in-law Dorothée Averkamp for the insightful discussions and constant mutual motivation that we shared while working virtually side by side. I am truly grateful for Tobias Bahn and Marc Sabek for our thoughtful discussions and the shared excitement for over a decade. I would like to genuinely thank Christian Busch, Timm Engelmeyer, Torben Engelmeyer and Jürgen Wicht for their support.

I would like to express my sincere appreciation to my supervisor, Uta Pigorsch, for her valuable guidance and support throughout my time as a PhD student. Moreover, I sincerely thank Florian Kluge who always helped me to tame tigers, defeat titans, face nemesis and to be the eliminator of all IT problems.

Furthermore, I want to express my gratitude to all the participants at the International Symposium on Forecasting 2020 hosted from Rio de Janeiro, Brasil and 2023 hosted in Charlottesville, United States of America. The discussions and helpful comments I received during the conferences were very useful for my research.

Contents

List of Figures

List of Tables

1 Introduction

Forecasts are an indispensable component of everyday life for individuals, society, economics, and businesses alike. Early humans had to observe and predict the weather to decide if conditions are suitable for hunting. On a larger scale, today's forecasts encompass traffic on both the road and regional levels to design and manage infrastructure, the energy consumption or energy generation by renewable sources to supply demand, or expected arrivals at emergency departments (Jiang & Luo, 2022; Petropoulos et al., 2022). On an economic level, forecasts for influential key figures such as the gross domestic product, the inflation rate, and the unemployment rate help to assess the current state of the economy and to inform decision-making. In financial applications forecasts for the volatility of returns are employed to assess risk and uncertainty. In food retail, companies forecast the daily demand of thousands of products for thousands of stores over multiple countries each day to ensure food supply (Petropoulos et al., 2022). There are numerous additional areas and applications where forecasts are utilized, including the introduction of new products, reverse logistics, interest and exchanges rates, stock returns, electricity prices, climate change, epidemics and pandemics, risk of violence, elections, or sports. For an extensive overview of examples, see Petropoulos et al. (2022).

In 1906 at the West of England Fat Stock and Poultry Exhibition a forecast of a different kind was required: the weight of a presented ox (after slaughtering and processing). It was a competition visitors could attend by paying a fee and that offered prices for the most accurate forecast. Among the 878 participants were butchers and farmers, i.e., people with experience and expert knowledge. Galton (1907) analyzed the forecasts of the participants and, after some initial errors in the analysis, they found that the average, i.e., a combination, of all forecasts is exactly equal to the weight of the ox, see also Wallis (2014); Wang, Hyndman, Li, and Kang (2023).

Although, this early observation hinted in the direction, the field of forecast combination was only formally defined and popularized many decades later by J. M. Bates and Granger (1969). They assess forecast combination as an optimization problem that minimizes the forecast error variance of the combined forecasts. Forecast combination is beneficial, because it mitigates sources of uncertainty that a single forecasting model (or expert) has: uncertainty of data, parameter, and model (expert). A forecast from a single model or expert is based on certain data or experience. However, other datasets or experiences can have additional valuable, independent information. This information is not ignored when combining forecast but leveraged. The estimated parameters of fore-

cast models have estimation uncertainty and one selected model can be less appropriate than others. Forecast combination mitigates those uncertainties and risks by diversification. A combined forecast is more robust and improves forecast accuracy (see e.g., J. M. Bates and Granger 1969; Clemen 1989; Newbold and Harvey 2008, pp. 268-269; Wang et al. 2023).

After the initial work by J. M. Bates and Granger (1969) many other methods have been developed ranging from simple combination schemes to more complicated and sophisticated approaches (Clemen, 1989; Wang et al., 2023). Additionally, multiple comparisons and competitions for time series forecasts have been conducted throughout the years. The most famous are the M competitions (see Makridakis et al., 1982, 1993; Makridakis & Hibon, 2000; Makridakis, Spiliotis, & Assimakopoulos, 2018, 2020, 2022; Makridakis et al., 2023). For a set of time series, participants provide forecasts that are then evaluated. These larger empirical comparisons have the intention to gain insights into the current state and future development of forecasting. Forecast combination has proved to be a competitive and viable approach that improves forecast accuracy (see also Bojer & Meldgaard, 2021). For example, in the M4 competition twelve of the 17 best performing method where combinations rather than single models (Makridakis et al., 2018, 2020).

However, throughout the last decade there was an early controversy and a puzzling observation or phenomenon around forecast combination. An early study by Newbold and Granger (1974) found that forecast combination improves forecast accuracy. However, there is a point of view that there is the one *true* model that describes a data generating process. The concept of forecast combination itself and the evidence that it can outperform traditional time series methods contradicts this view which led to some early controversy and heated discussion. Until today, it is *"a view commonly held [...] that there is some single model that describes the data generating process, and that the job of a forecaster is to find it. This seems patently absurd to me — real data comes from much more complicated, non-linear, non-stationary processes than any model we might dream up — and George Box himself famously dismissed it saying, 'All models are wrong but some are useful'."* (Hyndman, 2020), see also Box, Luceño, and Del Paniagua-Quinones (2009).[1]

This highlights and important premise for this thesis. Forecasting and forecast combination intends to provide the best possible forecast for given data. We do not seek to find one true model, method, or parameter setup that is superior to everything else always.

Beside the early controversy, there is a phenomenon surrounding forecast combination commonly referred to as the *"forecast combination puzzle"* (Stock & Watson, 2004). Of-

[1] In Hyndman (2020) the author (Editor-in-Chief of the Journal of International Forecasting from 2005-2018) reviews and discusses the history of forecast competitions.

tentimes, simple combination method, like the equally weighted forecasts that predicted the weight of the ox in 1906 England, outperform more sophisticated and theoretically superior methods. There is evidence that the forecast combination puzzle is caused by the estimation error of the weights assigned to each forecast that determine the combined forecast (Smith & Wallis, 2009).

In light of the aforementioned forecast combination puzzle and evidence indicating that it is caused by the estimation error of weights, the overarching research question for this thesis is: *how to further improve the forecast accuracy of a combined forecast using constrained weights?* Previously, we mentioned various application where forecasting and, thus, forecast combination is important. Each forecast application can have different requirements or characteristics of forecasts that are important, e.g., in food retail we can be interested in forecasts that are particularly well-suited for forecasting demand during a promotional period. This leads to our second research question: *how to incorporate additional, external information in forecast combination with constrained weights?*

In this thesis, we consider forecast combination with constrained weights, in particular shrinkage methods. To this end, we use the variance-minimization approach of the original forecast combination problem or approach by J. M. Bates and Granger (1969) and implement additional constraints. The constraints restrict weights and, thereby, shrink them towards a predefined direction depending on a shrinkage intensity or shrinkage parameter.

We analyze existing methods and propose extension to methods that shrink all weights simultaneously using an L_1 constraint in Chapter 4. To this end, we define a unified framework in form of an optimization problem. It is used to implement and compare the considered L_1 constraint approaches on the same basis and in the same form as the original forecast combination problem by J. M. Bates and Granger (1969) (see Diebold & Shin, 2019; Radchenko, Vasnev, & Wang, 2023; Roccazzella, Gambetti, & Vrins, 2022).

In forecast combination there are methods that impose a lower bound for weights. In Chapter 5 we propose to extend this idea by also implementing an upper bound. The proposed approach of *Forecast Combination with Bounded Weights (BW)* nests competitive benchmark methods for forecast combination, including the well-known and oftentimes superior equal weights forecast.

In Chapter 6 we propose a new direction: *Forecast Combination with Individual Feature Bounds (IFB)*. It does neither constraint all weights simultaneously (L_1 methods) nor commonly (Bounded Weights). While some combination methods use features or characteristics of the forecasts to estimate weights (see e.g., Kolassa, 2011), we use features or characteristics of the forecasts to define individual bounds for each weight. The more favorable the feature values of a forecast are, the less its weight will be constrained.

The IFB method allows incorporating additional, external information into the forecast combination. For some application the external feature can be the accuracy or diversity of the input forecasts, i.e., the forecasts that are combined. For other application it can be more specifically tailored to the application, e.g., financial key performance indicators or the accuracy and diversity of forecasts but for specific events like promotions in food retail.

In order to assess and compare the forecast combination methods, we will use both an extensive simulation study and a large scale empirical application. We adopt and extend a simulation study from Roccazzella et al. (2022). We use the simulation study to analyze the accuracy or performance of forecast combination methods for multiple different scenarios in a controlled environment. The empirical analysis provides a real-world comparison of the capabilities of the forecast combination methods. To this end, we use about 1000 monthly time series from the M4 competition to ensure meaningful and valuable analysis (Makridakis et al., 2018, 2020).

The main contributions of this thesis are:

(I) We provide an extended framework for simulation studies for forecast combination based on Roccazzella et al. (2022). To this end, we took inspiration from the well-known the Europeans Centrals Banks Survey of Professional Forecasters (Bowles et al., 2007; Garcia, 2003).

(II) We present a unified framework in form of an optimization problem. It incorporates the all L_1 constraint forecast combination methods and different shrinkage directions considered in this thesis.

(III) We propose to use what we call *Conditional Group Equal Weights (CGEW)* as a shrinkage direction for forecast combination with additional constraints.

(IV) We propose *Forecast Combination with Bounded Weights (BW)*: An extended approach for forecast combination that nests existing benchmarks, including equal weights. Thereby, Bounded Weights utilizes the advantages of the benchmark methods while mitigating their flaws.

(V) We propose a new direction for forecast combination: *Forecast Combination with Individual Feature Bounds (IFB)*. We implement individual bounds for each weight. The bounds are determined based on feature values or characteristics of the input forecasts.

(VI) We assess the forecast accuracy of the considered methods in an extensive simulation study and more importantly in a comprehensive empirical analysis.

The remainder of this thesis is organized as follows. In Chapter 2 we introduce the basic concepts of forecasting, forecast combination, discuss the forecast combination

puzzle and give a brief literature overview. In Chapter 3 we analyze forecasts from the Europeans Centrals Banks Survey of Professional Forecasters. Additionally, we present the framework for the simulation study that we use to assess the forecast accuracy of the considered forecast combination methods in the following three chapters. In Chapter 4, we present the idea and concept behind shrinkage, provide an overview of how the L_1 constraint is used for forecast combination so far and develop a unified framework for the L_1 constraint in form of an optimization problem. In Chapter 5 we incorporate an upper bound into the forecast combination optimization problem and analyze its effects. In Chapter 6 we present our new approach that introduces individual bounds that are determined based on feature values of the forecasts. In Chapter 7 we analyze the forecast performance of the considered forecast combination methods for about 1000 monthly time series. Chapter 8 briefly summarizes and discusses the content and result of this thesis.

2 Introduction to Forecast Combination

In this section we provide a brief introduction into the field of forecast combination. It is an integral part of forecasting literature and in 2021 roughly $13 - 14\%$ among published forecasting paper within the Web of Science concerned forecast combination (Wang et al., 2023). Throughout more than half a century, forecast combination proved itself to be a competitive approach that improves forecast accuracy. Both compared to the input forecast that are combined as well as other forecasting methods (see e.g., Bojer & Meldgaard, 2021; Hyndman, 2020; Makridakis et al., 2018, 2020). By combining multiple forecast to one, we use information from different data sets or expert knowledge and create are more accurate and robust forecast by diversification (see e.g., J. M. Bates and Granger 1969; Clemen 1989; Newbold and Harvey 2008, pp. 268-269; Wang et al. 2023).

In Section 2.1 we will discuss some basic concepts of forecasting that are important throughout this thesis. Thereafter, in Section 2.2, we introduce the forecast combination problem itself and analyze its solution. In Section 2.3 we will discuss the phenomenon that challenges the forecast combination community: sophisticated forecast combination method oftentimes have an inferior forecast accuracy compared to simple combination schemes. Lastly, we will provide a brief overview of a curated set forecast combination areas in Section 2.4.

2.1 The Basic Concepts of Forecasting

Given a time series $y_t \; \forall \; t = 1, \ldots, \tau$ the objective of forecasting is to predict or forecast future values of $y_t \; \forall \; t = \tau + 1, \ldots, T$ as accurate and precisely as possible. We denote forecasts for h-steps into the future by $\hat{y}_{t+h}$. However, note that in the literature forecast can also be written as $\hat{y}_{t+h|t}$ to indicate that it is based on the information up to time t. Within this thesis we will focus on one-step ahead forecasts, i.e., $h = 1$. The theory of time series forecasting is build around the premise that knowledge from historical patterns of y_t can be extended into the future. Additionally, exogenous variables can be used to explain y_t. In the field of forecasting there are multiple potential targets of a forecast. It can be the expected value, i.e., a point forecast, a prediction interval or the whole future distribution for the variable of interest (Petropoulos et al., 2022; Wang et al., 2023). Within this thesis we will focus on point forecasts.

Let there be multiple forecasts that can originate from different methods, algorithms, experts or are the result of forecast combination. Naturally, one needs to assess the value they provide and compare them with each other. To this end, we need a measurement to evaluate and compare forecasts and second we need to define a procedure on what basis we do so. There are multiple measurements that can be used which have different advantageous and disadvantageous. We will discuss those in more detail within the context of Chapter 6 in Section 6.1.3.1. For now, we will consider the *mean squared error (MSE)* because it is a prominent and oftentimes used measurement. It is based around the forecast error:

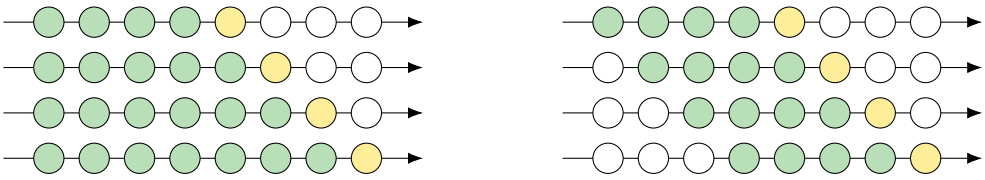
$$\varepsilon_t = y_t - \hat{y}_t. \tag{2.1}$$

The mean squared forecast error is then given by

$$\text{MSE} = \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} \varepsilon_t^2, \tag{2.2}$$

(see e.g., Hyndman & Koehler, 2006; Petropoulos et al., 2022; Thomson, Pollock, Önkal, & Gönül, 2019). Beside the choice of measurement, one needs future data to evaluate the forecasts on in order to assess the forecast accuracy. To this end, one can a procedure called pseudo out-of-sample forecasting. The available observations of the time series y_t $\forall\, t = 1, \ldots, T$ are split into a training set $t = 1 \ldots, \tau$ (in-sample) and a test set $t = \tau + 1, \ldots, T$ (out-of-sample). The training set is used to determine or train the forecasting method or algorithm and then forecasts are computed for the test set. Importantly, the test set has to be unknown to the method, i.e., no information from the test set can be used for training the method. As a result of pseudo out-of-sample forecasting one has $T - \tau$ forecasts and the true values of y_t to assess the forecast accuracy of the method. There are, however, several variants for pseudo out-of-sample forecasting. We want to emphasize two of them, the expanding and rolling window pseudo out-of-sample forecasting. Both expanding and rolling window have in common that they can be used to evaluate one-step ahead forecasts and that the method or algorithm can be re-estimated or re-trained multiple times.

Figure 2.1 illustrates the two variants. Each circle represents a time series observation. Green circles indicate that an observation is in the training set and a yellow circle is the currently considered observations in the test set. Observations depicted by a white circle are not taken into consideration at that point. The first four observations from left to right in both Figures 2.1(a) and 2.1(b) are the initial training set and the last four observations are the test set. The basic idea is that based on the training set y_t $t = 1, \ldots, \tau$ (green) one trains a method and produces a forecasts $\hat{y}_{t+1}$ for the first observation in the test set (yellow), see the first row in both Figures 2.1(a) and 2.1(b).



(a) Expanding window **(b)** Rolling window

Figure 2.1. Illustration of pseudo out-of-sample forecasting based on Hyndman and Athanasopoulos (2021).

After that the training set is updated or expanded by $\tau + 1$ and then based on the new training set one again trains a method, produces forecasts et cetera (second row). In case of the expanding window in Figure 2.1(a), the new training set contains all available observations for $t = 1, \ldots, \tau+1$. In case of a rolling or fixed window depicted in Figure 2.1(b) the first observations is omitted, i.e., the training set contains observations for $t = 2, \ldots, \tau+1$. To put it differently, the size of the training set constantly increases for the expanding window approach and is fixed in case of the rolling window approach. The former approach utilizes all available information. The latter captures more recent structures of the time series by omitting outdated observations and thus the method or algorithm are tailored towards those structures (see e.g., Inoue, Jin, & Rossi, 2017; Petropoulos et al., 2022; Tashman, 2000).

If a forecasting method or algorithm has additional parameters that define it, called hyperparameters, one can use time series cross-validation to determine them. Basically one applies pseudo out-of-sample forecasting, but only for the current training set. Regardless of whether one uses an expanding or rolling window approach. The following figure illustrates how cross-validation is used.

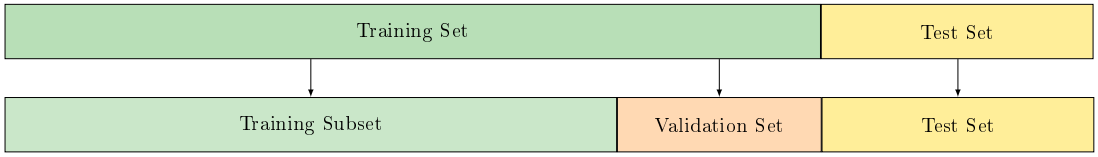


Figure 2.2. Illustration of pseudo out-of-sample forecasting with a validation set.

To this end, one still divides the data into a training set with observations $t = 1, \ldots, \tau$ depicted by the green rectangle in the first row in Figure 2.2. Te test set depicted by the yellow rectangle. Assume that we want to forecast the first observation in the test set $\tau+1$. To determine the hyperparameter we again divided the training set into a training subset and a validation set depicted in the second row as the green rectangle (less opacity) and the orange rectangle respectively. For each observation in the validation set we compute forecasts for candidate values of the hyperparameters using a rolling or expanding window. Then we choose the hyperparameters that minimize the measure of forecast accuracy for the validation set. Those hyperparameters are then used to

forecast the observation in the test set $\tau + 1$. The process is then repeated either by an expanding or rolling window (Hyndman and Athanasopoulos 2021; Inoue et al. 2017; G. James, Witten, Hastie, Tibshirani, and Taylor 2023, pp. 201-208; Petropoulos et al. 2022; Tashman 2000).

2.2 The Basic Concepts of Forecast Combination

In this section we will introduce the work of J. M. Bates and Granger (1969) in Section 2.2.1. They popularized forecast combination and defined the forecast combination problem for two forecasts. Thereafter, in Section 2.2.2 we present the generalization of the forecast combination problem that can be used for an arbitrary number of forecasts. Lastly, in Section 2.2.3 we take a closer look at the solution or weights of the forecast combination problem, especially negative weights.

2.2.1 Forecast Combination with Two Forecasts

The research area of forecast combination follows from the initial work by J. M. Bates and Granger (1969). The main objective of forecast combination is to combine a set or pool of forecasts such that the combined forecasts has a superior out-of-sample forecast accuracy (MSE) compared to the forecasts that are combined. To this end, J. M. Bates and Granger (1969) proposed to combine two forecasts such that the in-sample *error variance* (≈MSE if the forecast error is unbiased) is minimized.

Assume a time series y_t with an expected value of μ and two unbiased one-step ahead forecasts for y_t, i.e., $\hat{y}_{1,t}$ and $\hat{y}_{2,t}$ with $E[\hat{y}_{1,t}] = \mu$ and $E[\hat{y}_{2,t}] = \mu$. J. M. Bates and Granger (1969) define the combined forecast for $t + 1$, i.e., $\hat{y}_{c,t+1}$, by

$$\hat{y}_{c,t+1} = \omega_1 \hat{y}_{1,t+1} + \omega_2 \hat{y}_{2,t+1}. \tag{2.3}$$

with $\omega_2 = (1 - \omega_1)$. For the sake of simplicity we follow Radchenko et al. (2023) and will, henceforth, denote any forecast i, that can originate from different methods, algorithms or experts, by $\hat{y}_i$ as an abbreviation of $\hat{y}_{i,t+1}$. The combined forecast is given by $\hat{y}_c$ instead of $\hat{y}_{c,t+1}$ (J. M. Bates & Granger, 1969; Radchenko et al., 2023; Wang, Kang, & Li, 2022).

The combined forecast in Equation (2.3), $\hat{y}_c$, is a linear combination or weighted average of the two original forecast. The weights for the forecasts are denoted by ω_1 and ω_2. The expected value of the combined forecasts is

$$E[\hat{y}_c] = E[\omega_1 \hat{y}_1 + \omega_2 \hat{y}_2], \tag{2.4}$$

$$= \omega_1 E[\hat{y}_1] + \omega_2 E[\hat{y}_2]. \tag{2.5}$$

If the assumption of unbiased forecasts holds, it follows that $\omega_2 = 1 - \omega_1$ in order to ensure that the combined forecasts is also unbiased. Accordingly,

$$E[\hat{y}_c] = \omega_1\mu + (1-\omega_1)\mu, \tag{2.6}$$
$$= \mu = E[y]. \tag{2.7}$$

The constraint that weights have to sum up to unity is the *unity constraint* (J. M. Bates & Granger, 1969; Granger & Ramanathan, 1984; Radchenko et al., 2023). In this thesis we will focus on forecast combination methods that fulfill the unity constraint.

In order to calculate the combined forecasts in Equation (2.3), weights have to be determined. Recall that the objective of the forecast combination problem from J. M. Bates and Granger (1969) is to minimize the in-sample error variance. It is based on the forecast error first introduced in Equation (2.1). The forecast error, ε_c, of the combined forecast of Equation (2.3) is given by

$$\varepsilon_c = y - \hat{y}_c, \tag{2.8}$$
$$= y - \big(\omega_1\hat{y}_1 + (1-\omega_1)\hat{y}_2\big), \tag{2.9}$$
$$= y - \omega_1\hat{y}_1 - \hat{y}_2 + \omega_1\hat{y}_2, \tag{2.10}$$
$$= y - \omega_1\hat{y}_1 - \hat{y}_2 + \omega_1\hat{y}_2 + \omega_1 y - \omega_1 y, \tag{2.11}$$
$$= \omega_1(y - \hat{y}_1) + (1-\omega_1)(y - \hat{y}_2), \tag{2.12}$$
$$= \omega_1\varepsilon_1 + (1-\omega_1)\varepsilon_2. \tag{2.13}$$

Thus, the forecast error of the combined forecast is a linear combination of the forecast errors ε_1 and ε_2 of the two original forecast $\hat{y}_1$ and $\hat{y}_2$. Before we derive the error variance of the combined forecast, note that the error variance of any forecast $\hat{y}_i$ is defined as $Var(\varepsilon_i) = \sigma_i^2$ and its standard deviation is σ_i. Furthermore, $\sigma_{i,j}$ and $\rho_{i,j} = {}^{\sigma_{i,j}}/_{\sigma_i\sigma_j}$ are the covariance and correlation respectively between the forecast errors ε_i and $\varepsilon_j \; \forall \; i, j = 1\ldots, N$ (see e.g., Fahrmeir, Heumann, Künstler, Pigeot, & Tutz, 2016, pp. 323-330). It is important to notice that those measures are based on the forecast error, i.e., we consider the *error variance*, *error covariances* and *error correlations*. Therefore, the error variance of the combined forecast σ_c^2 is given by

$$\sigma_c^2 = E[\varepsilon_c^2] \tag{2.14}$$
$$= E[\big(\omega_1\varepsilon_1 + (1-\omega_1)\varepsilon_2\big)^2], \tag{2.15}$$
$$= \omega_1^2 E[\varepsilon_1^2] + (1-\omega_1)^2 E[\varepsilon_2^2] + 2\omega_1(1-\omega_1)E[\varepsilon_1\varepsilon_2], \tag{2.16}$$
$$= \omega_1^2\sigma_1^2 + (1-\omega_1)^2\sigma_2^2 + 2\omega_1(1-\omega_1)\sigma_{1,2}, \tag{2.17}$$
$$= \omega_1^2\sigma_1^2 + (1-\omega_1)^2\sigma_2^2 + 2\rho_{1,2}\omega_1\sigma_1(1-\omega_1)\sigma_2. \tag{2.18}$$

The error variance from Equations (2.17) and (2.18) is the objective function of the forecast combination problem. Note that the more common representation of the combined error variance is given in Equation (2.17), however, Equation (2.18) is more intuitive. Therefore, for now we will focus on Equation (2.18). The error variance of the combined forecast is determined by the forecast error variances of the two considered forecasts σ_1^2 and σ_2^2 and by their corresponding standard deviations σ_1 and σ_2. Lastly, the error correlation $\rho_{1,2}$ between the two forecast errors ε_1 and ε_2 is part of σ_c^2. The error correlation indicates the relationship between the forecasts errors. Note that Equation (2.17) this relationship is considered by the error covariance, i.e., the not standardized error correlation. A positive error correlation implies that the forecasts tend to make similar errors. The degree of similarity is indicated by the error correlation coefficient. In summary, the error variance of the combined forecasts is influenced by the forecasts accuracies as well as the relationship between the forecasts errors (see e.g., J. M. Bates and Granger 1969; Elliott and Timmermann 2016, pp. 313-315; Newbold and Harvey 2008, pp. 270-271). We will revisit the components of the combined error variances throughout this thesis in form of Equation (2.17).

Based on the objective function, i.e., the error variance of the combined forecast, from Equation (2.17), one can derive the optimal weights that minimize it. Those weights, i.e., the *optimal weight (OW)*, are given by

$$\omega_1^{OW} = \frac{\sigma_2^2 - \rho_{1,2}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho_{1,2}\sigma_1\sigma_2}., \tag{2.19}$$

(J. M. Bates & Granger, 1969; Smith & Wallis, 2009). If the optimal weights are used to combine the forecast, it holds that

$$\sigma_c^2 \leq \min\left\{\sigma_1^2, \sigma_2^2\right\}. \tag{2.20}$$

The result shown in Equation (2.20) is major cornerstone of forecast combination. It means, that the in-sample error variance of the combined forecast is at most equal to or lower than the smallest of the forecasts error variances. To put it differently, combining forecasts is never inferior compared to the forecast that are combined. (J. M. Bates & Granger, 1969; Dickinson, 1975).

Before we move on we have to clarify a potential ambiguity regarding the forecast error, error variance, and MSE. We want to emphasize that the in-sample data for forecast combination are forecasts. Although forecast errors refer to out-of-sample forecast, within forecast combination, they are the in-sample difference between the actual series and the candidate, input forecasts for combination. Based on the in-sample forecast errors one calculates the error variance and covariances which are used to determine weights. These weights minimize the in-sample error variance. The MSE of the in-sample forecast errors is in theory identical to the error variance depicted in

Equations (2.2) and (2.8) if forecast errors are unbiased. In practice, because the mean of forecast errors in not exactly zero, the MSE is an approximation of the error variance (Granger & Ramanathan, 1984; Wang et al., 2023) or, in other words, it is the empirical error variance of the input forecasts. The relationship of error variance and MSE is crucial to keep in mind throughout this thesis. Particularly in the context of the simulation study we use the term *error variance* which, again, is a measure for the forecast accuracy of the input forecasts.

2.2.2 Forecast Combination with N forecasts

J. M. Bates and Granger (1969) introduced forecast combination for $N = 2$ forecasts. To adapt it for an arbitrary number of forecasts $N \in \mathbb{N}_{\geq 2}$, the combined forecast is defined as

$$\hat{y}_c = \hat{\boldsymbol{y}}'\boldsymbol{\omega} \tag{2.21}$$

with the $N \times 1$ vector $\hat{\boldsymbol{y}}$ that contains all forecasts $\hat{y}_i \ \forall i = 1, \ldots, N$ and the $N \times 1$ vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)'$ contains the weight for each forecast. The *forecast combination problem* is given by

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1 \end{aligned} \tag{2.22}$$

(see e.g., Elliott & Timmermann, 2016, pp. 313-315). Note that the forecast combination problem present in Equation (2.22) is the basis of the combination approaches considered in this series and will be referenced repeatedly throughout it. The objective function consists of the weight vector $\boldsymbol{\omega}$ and the $N \times N$ matrix $\boldsymbol{\Sigma}$. The latter is the variance-covariance matrix of the forecast errors, i.e.,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,N} \\ \sigma_{1,2} & \sigma_2^2 & \ldots & \sigma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,N} & \sigma_{2,N} & \ldots & \sigma_N^2 \end{pmatrix}. \tag{2.23}$$

Accordingly, the objective function can be written as

$$\boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{i,j} \tag{2.24}$$

where $\sigma_{i,i} = \sigma_i^2$. For $N = 2$, this is identical to the in-sample error variance of the combined forecast σ_c^2 given in Equation (2.17). Accordingly, the objective function is to minimize the error variance of the combined forecast. The single constraint within

the optimization problem of Equation (2.22) is the already described unity constraint, i.e., all weights sum up to one. To impose it, we multiply the transpose of $\boldsymbol{\omega}$ by a $N \times 1$ vector $\mathbf{1}$ that contains only ones. The optimal weights, i.e., solution $\boldsymbol{\omega}^{OW}$ to the optimization problem, are

$$\boldsymbol{\omega}^{OW} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\imath}}{\boldsymbol{\imath}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\imath}}. \tag{2.25}$$

In the same sense as for $N = 2$, recall Equation (2.20), with $\boldsymbol{\omega}^{OW}$ for the combined forecast it holds that

$$\sigma_c^2 \leq \min\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2\}, \tag{2.26}$$

i.e., it can result in a better error variance than all forecasts and in the worst case its error variance is identical to that of the best forecast (Elliott and Timmermann 2016, pp. 313-315; Newbold and Harvey 2008, pp. 270-271, Dickinson 1975). The optimal weights $\boldsymbol{\omega}^{OW}$ minimize the in-sample error variance σ_c^2 of the combined forecast. Therefore, we will refer to this approach introduced by J. M. Bates and Granger (1969) as the *optimal weights (OW)* approach. However, it is important to emphasize that the input in the forecast combination problem of Equation (2.22) is the true error variance-covariance matrix $\boldsymbol{\Sigma}$. In reality, it is unknown and the empirical error variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$ needs to be estimated and then used to determine weights by Equation (2.25) (Radchenko et al., 2023).[2] We will further discuss this in Section 2.3.

Based on the MSE loss, Granger and Ramanathan (1984) showed that regression models can also be used to combine N forecasts. Each forecast is a regressor and the actual value y is the dependent variable, i.e.,

$$y_t = \omega_0 + \sum_{i=1}^{N} \omega_i \hat{y}_{i,t} + \epsilon_i, \tag{2.27}$$

where ϵ_i is the regression error term. Weights are then estimated based on the ordinary least squares (OLS) estimator which minimizes the squared residuals. They used regression models both with and without the intercept ω_0. They also include a model without an intercept but with the additional constraint that weights have to sum up to one, i.e., a unity constraint. The solution of this regression model is an approximation of the optimal weights (OW) approach (Granger & Ramanathan, 1984; Wang et al., 2023).

[2] Note that within this thesis we use the simple sample error variance covariance estimator. See for example the sample variance as a special case of the covariance in Fahrmeir et al. (2016, pp. 64-65).

2.2.3 Positive and Negative Weights

In this section we take a closer look at the forecast combination weights. For the sake of simplicity we will consider two forecasts $\hat{y}_1$ and $\hat{y}_2$. Recall that the weight ω_1 for the first forecast is determined by Equation (2.19). The weight for the second forecast is $\omega_2 = (1 - \omega_1)$. For now assume that the forecast errors are uncorrelated, i.e., $\sigma_{1,2} = \rho_{1,2} = 0$. As a result ω_1 is defined by

$$\omega_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{2.28}$$

If both forecasts have the same error variance, the optimal weights are equal. This solution is called *equal weights (EW)* and it will be important throughout this thesis. Again, it is given if the forecast errors have the *identical error variance* and are *uncorrelated*. Now assume without loss of generality that $\sigma_1^2 < \sigma_2^2$, i.e., the first forecast has a smaller error variance. From Equation (2.28) it follows that its weights will be larger, i.e., $\omega_1 > \omega_2$. The larger the difference between the error variance is, the closer the weight of the first forecast is to one. It holds that ω_1 is only equal to one, if its error variance is zero, i.e., it perfectly forecasts future values of y. Accordingly, if forecast errors are uncorrelated, for both weights it holds $\omega_i \in [0, 1] \; \forall \; i = 1, 2$. Note that J. M. Bates and Granger (1969) even presented Equation (2.28) as a more simple method to combine forecasts, although forecast error are correlated (Radchenko et al., 2023; Smith & Wallis, 2009; Winkler & Clemen, 1992).

If the forecast error correlation is taken into consideration, it holds that $\omega_i \in \mathbb{R}$. To put it differently, weights outside the interval of $[0, 1]$ can occur, i.e., both negative weights and weights greater one. In case of two forecasts, negative weights and weights greater one only occur simultaneously. Otherwise, the unity constraint is violated. In general ($N > 2$), there can be negative weights without weights greater one, e.g., $\omega_1 = -0.2$, $\omega_2 = \omega_3 = 0.6$. The opposite is not true, because weights greater one require negative weights to fulfill the unity constraint. Overall a situation where weights are outside the $[0, 1]$ interval is usually discussed in terms of negative weights in the literature (see e.g., Radchenko et al., 2023). To put it differently, if the forecast error correlation is taken into consideration, the absolute sum of weights can be greater one. In conjunction with the unity constraint this is only feasible if negative weights occur. J. M. Bates and Granger (1969) observed the occurrence of negative weights and provide an example to get an intuition for it. Assume that there are two forecasts $\hat{y}_1 = 80$ and $\hat{y}_2 = 100$. If the true value $y = 120$, the only way that the combined forecasts is $y_c = 120$ is to use $\omega_1 = -1$ and $\omega_2 = 2$, i.e., negative weights.

Negative weights are possible if the error correlations are taken into consideration. Radchenko et al. (2023); Winkler and Clemen (1992) analyzed the solution of the forecast combination problem and found that the larger the difference or ratio between the

error variances as well the higher the correlation, the more likely negative weights occur. In fact, negative weights can only occur for positive correlations as one can see by considering Equation (2.19). The weight of the first forecast ω_1 is negative if $\sigma_2^2 < \rho_{1,2}\sigma_1\sigma_2$. Because $\sigma_1, \sigma_2 > 0$ by definition, it only occurs for positively correlated forecast errors. Against the background of the example of J. M. Bates and Granger (1969) discussed previously this is sensible. Only if the forecast errors are positively correlated there are likely to both over- or underestimated the true value simultaneously.

Without loss of generality assume that $\sigma_i^2 < \sigma_j^2$. Winkler and Clemen (1992) showed that negative weights occur if

$$\sigma_i^2/\sigma_j^2 > \rho_{i,j}. \tag{2.29}$$

Accordingly, more dissimilar error variance lead to negative weights for smaller error correlations. Moreover, note that the higher the correlation, the more sensitive the weights are to small changes in the error variances, i.e., σ_i^2/σ_j^2 (Radchenko et al., 2023; Winkler & Clemen, 1992).

Consider now the forecast combination problem of Equation (2.22) with $N > 2$ as analyzed by Radchenko et al. (2023). They derived conditions under which negative weights occur. Overall, they show that negative weights occur for $N > 2$ if forecasts errors are highly positively correlated (for more details see Radchenko et al., 2023).

Although negative weights can occur in the optimal solution, i.e., minimize the error variance of the combined forecast, there are many methods that suggest to only consider positive weighs. Even J. M. Bates and Granger (1969) suggested omitting the information of the error correlation and use Equation (2.28) to determine weights for forecast combination. Further research supported this strategy (see e.g Clemen, 1989; Smith & Wallis, 2009). For example, Aksu and Gunter (1992) and Gunter (1992) added a non-negativity constraint to the regression model of Equation (2.27), i.e., a feasible solution has to consist of only positive weights including zero. Conflitti, de Mol, and Giannone (2015) included the non-negativity constraint into the optimization problem of Equation (2.22). The corresponding problem is given by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{w}'\mathbf{1} = 1 \\ & \omega_i \geq 0 \qquad \forall\, i = 1, \ldots, N. \end{aligned} \tag{2.30}$$

We follow Nowotarski, Raviv, Trück, and Weron (2014) who used this approach within an empirical study and refer to it as *positive weights (PW)*. Henceforth, we will use the PW approach in form of Equation (2.30) as a benchmark throughout this thesis, due to its promising results in previous studies (Conflitti et al., 2015; Nowotarski et al., 2014; Wang et al., 2023). Nevertheless, an advantage of negative weights is that

they can correct the combined forecast if the input forecasts simultaneously over- or underestimate the true value (Roccazzella et al., 2022) Therefore, we will not rule out negative weights but look at a way to constrain them in Chapters 4 to 6. However, before we move on we have to discuss why such constraints are necessary. Throughout the last half century forecast combination has in fact proven itself to be a very powerful tool, however, not when using the OW approach from J. M. Bates and Granger (1969).

2.3 The Forecast Combination Puzzle

The optimal weights from the forecast combination problem by J. M. Bates and Granger (1969) presented in Equation (2.22) is the dominant strategy to combine forecasts, however, only if the true error variance covariance matrix is known. Then, the error variance of the combined forecast is at least as good as the smallest error variance among all forecasts that are used as input. However, over the last half century a curious observation puzzles the field of forecasting combination: simple combination methods outperform more sophisticated approaches, including the OW approach. Especially, the simple average of forecasts, or equal weights forecast as it is more commonly referred to, is a tough benchmark to beat (see e.g., Genre, Kenny, Meyler, & Timmermann, 2013; Wang et al., 2023). This phenomenon was addressed early on in Clemen (1989) who provided an annotated bibliography on different forecast combination approaches. They suggested that further research is needed to analyze why the simple average is oftentimes the best or close to the best performing method. Stock and Watson (2004) analyzed forecasting output growth. Their results are in line with empirical evidence that showed the equal weights forecast to have a superior forecast accuracy compared sophisticated methods. They referred to this phenomenon as the "*forecast combination puzzle*" (Stock & Watson, 2004).

Early on, a potential explanation for the forecast combination puzzle was discussed in Clemen (1986). As shown by Granger and Ramanathan (1984) the forecast combination problem can be expressed as a regression model where the forecasts are used as variables. Forecast are oftentimes highly correlated and, by that, the regression model or forecast combination problem suffers from imperfect multicollinearity if the ordinary least squares estimator is used. As a result the estimated weights will have a large variance, i.e., a larger estimation error. For example, let there be $N = 2$ forecasts i, j that are combined by a linear regression model. The variance of the estimated weight of forecast i is given by

$$Var(\hat{\omega}_i) = \frac{1}{N}\left(\frac{1}{1-\rho_{i,j}^2}\right)\frac{Var(\varepsilon_c)}{Var(\hat{y}_1)}, \tag{2.31}$$

(Radchenko et al., 2023). The variance of the estimated weights becomes larger, the higher the correlation of the individual forecasts is. Recall, that the input of the forecast combination problem is the estimated variance-covariance matrix $\widehat{\mathbf{\Sigma}}$ which inhibits a certain estimation error. This in conjunction with the sensitivity of weights due to a large variance of them is a potential reason for the forecast combination puzzle (Claeskens, Magnus, Vasnev, & Wang, 2016; Clemen, 1986; Radchenko et al., 2023; Smith & Wallis, 2009).

Smith and Wallis (2009) provide evidence that the forecast combination puzzle is caused by the finite-sample estimation error of weights. Accordingly, the benefit of estimating weights in, for example, the optimal weights approach is exceeded or surpassed by the estimation error. Claeskens et al. (2016) show that the combined forecast is biased and has a larger variance if the weights are estimated. The result from Chan and Pauwels (2018) provides further evidence for the estimation error to be the cause of the forecast combination problem. Accordingly, the superiority of equal weights is based on the fact that weights do not need to be estimated (see also Wang et al., 2023).

In summary, the forecast combination puzzle seems to be due to several reasons: the estimation error of the weights in finite samples, the estimation error in the *estimated* variance-covariance matrix and the sensitivity of weights due to highly correlated forecast errors (Chan & Pauwels, 2018; Claeskens et al., 2016; Clemen, 1986; Smith & Wallis, 2009; Wang et al., 2023). The fact that the optimal weights approach does not have a superior performance empirically but succumbs simple combination methods like the equal weights forecast has sparked interest into developing further methods to combine forecasts or determine weights.

2.4 Brief Overview of Forecast Combination Methods

This section provides a brief overview of forecast combination methods that have been developed. Our literature review is based on Wang et al. (2023) who provided a review on the occasion that forecast combination was introduced more than half a century ago. For an extensive overview as well as reference to further readings, we therefore refer the reader to Wang et al. (2023).

After the initial work of J. M. Bates and Granger (1969) many more forecast combination methods have been developed. They mostly concern point forecasts, however, more recently the combination of probabilistic forecasts receives more attention (see e.g., Hall & Mitchell, 2007; Martin, Loaiza-Maya, Maneesoonthorn, Frazier, & Ramírez-Hassan, 2022; Wang et al., 2023). Recall, that within this thesis we focus on point forecasts.

Although the forecast combination puzzle is based on the surprisingly robust and superior forecast accuracy of the equal weights forecast, there are additional simple combination schemes that have been used. For example, the median, a trimmed, winsorized or

bias adjusted mean (Capistrán & Timmermann, 2009; Jose & Winkler, 2008). In comparison to the equal weights forecast, the median forecast is more robust to outliers. Two combination schemes that used the median (Petropoulos & Svetunkov, 2020) and mean (Shaub, 2020) of a pool of (simple) forecasts have recently proved their competitiveness to more sophisticated forecasting models in the M4 competition (Makridakis et al., 2020). However, there is no conclusive evidence against or for a general superiority of the mean or median (Kolassa, 2011; Wang et al., 2023).

The objective function of the OW approach by J. M. Bates and Granger (1969) presented in Equation (2.22) is to minimize the error variance, i.e., it is based on a symmetric (MSE) loss. Different, asymmetric and skewed loss function for the forecast combination problem are also taken into consideration in the literature. For example, Elliott and Timmermann (2004) showed that the optimal weights under the MSE loss are optimal for a variety of loss functions. However, it depends on how skewed the forecast errors are and on how asymmetrical the loss function is (Elliott and Timmermann 2016, pp. 313-315; Elliott and Timmermann 2004; Wang et al. 2023).

Other methods shrink weights towards reference values to improve the out-of-sample forecast accuracy of the combined forecast. Note that these methods are the focus of this thesis and will be discussed in detail in Chapters 4 to 6. Nevertheless, we continue with a brief overview. Diebold and Shin (2019) used variants of the *"least absolute shrinkage and selection operator" (Lasso)* introduced in Tibshirani (1996) for forecast combination. For the Lasso, the sum of absolute weights is incorporated into the objective function and, thus, weights are shrunken and selected towards zero. Diebold and Shin (2019) introduced the *egalitarian Lasso (eLasso)* that, in contrast, shrinks weights towards equal weights. They also presented a two-step procedure, the *partially egalitarian Lasso (peLasso)*, that first selects weights using the standard Lasso and then applies the eLasso. Radchenko et al. (2023) and Roccazzella et al. (2022) also work with Lasso-based approaches that additionally include the unity constraint.

Beside the optimal weights approach, J. M. Bates and Granger (1969) also discussed performance based weighting schemes. To this end, weights are determined by their inverse performance relative to all others. They suggested five such weighting schemes which partly omitted the error correlation or put more emphasize on more recent forecast errors. In a similar fashion information criterion, like the Akaike, corrected Akaike or the Bayesian information criterion can be used to determine weights (Kolassa, 2011; Wang et al., 2023).

In contrast to the frequentist approach of combining weights, research is also directed towards a Bayesian alternative. It allows to incorporate prior information into the estimation of combination weights (see e.g., Clemen, 1986; Diebold & Pauly, 1990; Wang et al., 2023)

In this thesis we consider linear combination approaches for a single time series. Nevertheless, we briefly introduce non-linear combination schemes and cross-learning approaches. Beside linear combination approaches, there is limited research in the area of non-linear combination schemes. To this end, input forecasts are combined by non-linear functions instead of linear ones. For example, Babikir and Mwambi (2016) and Krasnopolsky and Lin (2012) used artificial neural networks to combine forecasts. The results provide evidence for an improvement in forecast accuracy when using non-linear combination methods. However, more research is needed in this area (for more details see Wang et al., 2023). Another promising field of forecast combination is based on meta-learning or cross-learning approaches. To this end, combination weights are learned across different time series using machine learning algorithms. For example Montero-Manso, Athanasopoulos, Hyndman, and Talagala (2020) proposed an automated method FFORMA (Feature-based Forecast Model Averaging) that uses time series characteristics to determine weights to combine forecasts. FFORMA was the second most accurate contribution to the M4 Competition (Makridakis et al., 2018, 2020) both for the point forecasts and prediction intervals (Montero-Manso et al., 2020; Wang et al., 2023). We will discuss FFORMA in more detail within Chapter 6 (for more detail see also Wang et al., 2023).

Lastly we want to briefly discuss research that is directed towards selecting a subset of a pool of forecast instead of using all of them. To improve the forecast accuracy of a combined forecast. The simplest variable or subset selection technique is to only use the forecast with the best forecast accuracy(see e.g., Mannes, Soll, & Larrick, 2014). Alternatively, one can also look at other measures or features of forecast like their diversity to determine subsets (see e.g., Thomson et al., 2019). The lasso based methods used in Diebold and Shin (2019); Radchenko et al. (2023) and Roccazzella et al. (2022) that focus on shrinkage of weights can also incorporate a variable selection as we will discuss in Chapter 4. Prior to that, we first introduce a simulation study in the following chapter. We will use it both for illustrating the upcoming forecast combination methods and evaluate them.

3 Survey of Professional Forecasters and Simulation Study

In this chapter we design our simulation study and present the scenarios that we consider throughout this thesis. To this end, we first gather information from a real world example, the European Central Banks (ECB) *Survey of Professional Forecasters (SPF)*.[3] For the analysis of the SPF we are in particular interested in two aspects. First, the forecast accuracy, i.e., error variance, of the individual forecasts. Second, the correlation of the forecast errors. These components define the error variance covariance matrix, which is the core part of the objective function of the forecast combination problem. The SPF is a well-known data set for forecast combination, and it provides numerous expert forecasts (see e.g., Capistrán & Timmermann, 2009; Genre et al., 2013; Radchenko et al., 2023; Roccazzella et al., 2022). Our main focus with respect to the SPF is to gain insights from real-world data in order to create scenarios for our simulation study. To this end, in Section 3.1 we introduce the SPF, prepare the data and analyze the resulting set of forecasts. Based on that, in Section 3.2 we present a framework that can be used to design simulation studies for forecast combination. Additionally, we discuss the scenarios that we use throughout this thesis.

3.1 ECB's Survey of Professional Forecasters

The SPF and similar surveys are suited to be used for forecast combination, because external forecasts are provided from experts that can result from different data sets, knowledge and / or models (see e.g., Capistrán & Timmermann, 2009; Genre et al., 2013; Radchenko et al., 2023). In this thesis we will use the Survey of Professional Forecasters (SPF) from the European Central Bank (ECB) as an empirical, real world reference point upon which we build and expand our simulation study. For that purpose, this section provides a short overview of the operating principle, data preparation as well as a descriptive analysis of the ECB's Survey of Professional Forecasters. For a more detailed explanation of the SPF see Bowles et al. (2007); Garcia (2003).

The ECB's Survey of Professional Forecasters collects forecasts of experts regarding macroeconomic key figures since 1999. Experts are chosen based on their macroeconomic and forecasting expertise with special attention to the euro area. They are affili-

[3] The data for the SPF was downloaded from the website of the ECB `https://www.ecb.europa.eu` on the 16th November 2021.

ated to different institutions within the European Union from both the financial sector, e.g., banks, and non-financial sector, e.g., research institutions. To ensure independent forecasts, each expert or institution should not be closely connected to another expert or institution. The survey started with 95 experts and over the year more experts joined the survey leading to a number of 130.[4]

The ECB's Survey of Professional Forecasters is conducted on a quarterly base by sending questionnaires to the experts. The questionnaires are sent when the latest data of the macroeconomic key figures is available. This does not include data from the quarter itself at which the questionnaire send but only from previous quarters. In each questionnaire the experts are asked for point and density forecasts for certain macroeconomic key figures of the euro area.

The first key figure of interest is the *Harmonised Index of Consumer Prices (HICP) inflation rate*. It has a monthly frequency and is measured as the annual percentage change. Second, the monthly *unemployment rate* as a percentage of the labor force is considered. The third and last key figure is the growth rate of the *real Gross Domestic Product (GDP)*. Its frequency is quarterly, and it is measured as an annual percentage change. The overarching goal of the analysis of the SPF is to gain some insights into how forecast errors are correlated and how the forecast accuracy is distributed. It is not to fully analyze the SPF. Therefore, henceforth, we will focus solely on the HICP.

Although, for each macroeconomic key figure, five forecast horizons are requested from each expert, we will again focus on one specific horizon: the one-year ahead forecast. This was, inter alia, also considered in the related literature, see e.g., Genre et al. (2013); Matsypura, Thompson, and Vasnev (2018); Radchenko et al. (2023). For the HICP Inflation rate forecasts are for a specific month. The target month is the last month of the corresponding one-year ahead quarter. For example, for the SPF conducted in the first quarter of a year t, i.e., Q1, the latest monthly data is from December of year $t-1$. The forecasts are required for the target month of December in year t.

The first SPF was published in 1999Q1 and, at the time the data was downloaded, the latest available SPF is from 2021Q3. The first one-year-ahead forecast of the SPF is for December 1999 and, thus, our data set for the one-year-ahead forecast consists of 88 observations (points in time) (Bowles et al., 2007; Garcia, 2003).

3.1.1 Missing Observations in the Data Set

Missing values are a problem for forecast combination. First, we need a sufficiently large data set either to be able to estimate weights and furthermore to ensure that the weights can be estimated sufficiently precise. Second, for forecast methods like the approach by J. M. Bates and Granger (1969) we need to estimate the covariance matrix. However, in order to estimate the common sample covariance matrix, the data can not have missing

[4] The latest data within our data set is from the third quarter of 2021.

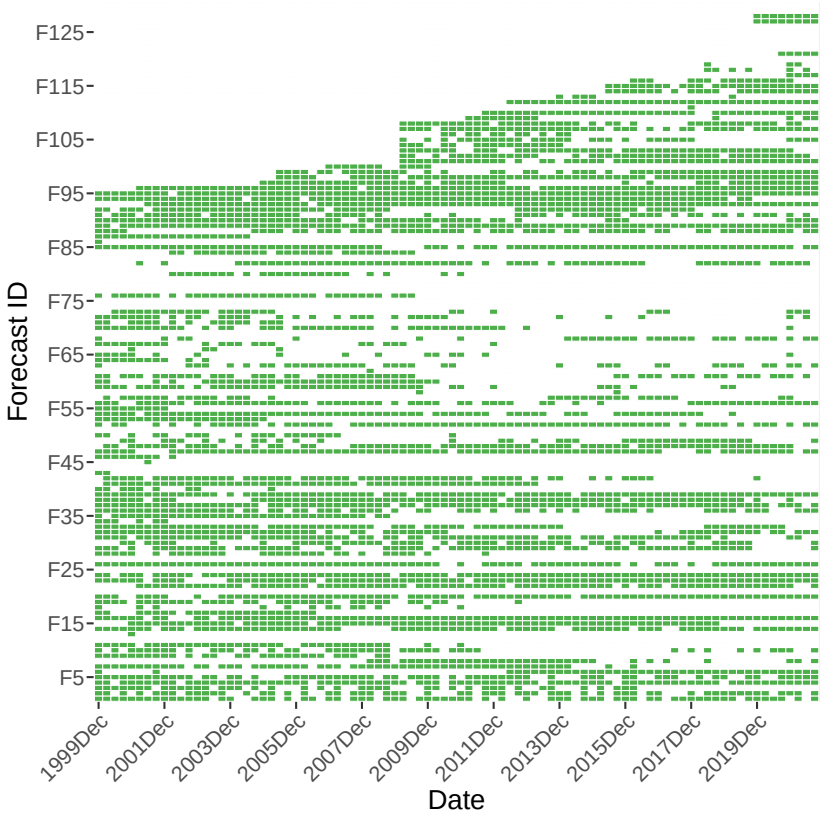


Figure 3.1. Visualization of expert responses to the SPF. A green rectangle depicts a response of the expert. The illustration is inspired by Capistrán and Timmermann (2009); Genre et al. (2013).

values.[5] However, within the SPF there are missing values, i.e., it is an unbalanced panel. The experts responses for the one-year-ahead forecast of the HICP inflation are depicted in Figure 3.1. The abscissa shows the date or, to be more precise, the target month for which the forecast was determined. The Forecaster ID is on the ordinate. A green rectangle indicates that the corresponding expert has provided a forecast for this target month. A white area indicates missing values. A brief contemplation of Figure 3.1 reveals immediately that missing values are prevalent in the survey.

Taking a more specific look, first and foremost, reveals that there are 22 out of 130 Forecaster that never provide inflation forecasts. Second, there are some experts that began responding rather frequently but stopped at some point. For example $F76$ did not contribute any forecasts after June 2009. Similarly, some experts joined the SPF later on and, of course, only submitted forecasts thereafter, for example $F105$. Third, even experts that respond frequently can have missing values, see for example $F4$. Overall,

[5] This hold similarly for the forecast combination regression framework presented Granger and Ramanathan (1984) in order to determine the OLS estimator.

the average response rate of the expert is roughly 56%. Noteworthy, for the whole data set there is not a single expert that is always responding.[6]

In our analysis, we use the rolling window approach, described in Section 2.1, with 44 observations, i.e., eleven years. Although, this increases the chance that we have experts that always respond within the currently considered subset of observations, ten of the subsets have not one consistently responding expert and another five only have one expert. The remaining subsets have between two and five always responding experts. Even under this rolling window approach we face the problem of missing values which we address in the following.

3.1.2 Filtering and Balancing the Data Set

In order to process the data such that it can be used for forecast combination we need to consider how to filter and balance the panel. It means we both select a subset from all available experts (filter) and then consider how to treat missing values, for example by balancing the panel. Note that, although we only use SPF to analyze the structure of forecast errors and accuracy, it is necessary to prepare it as it would be used for forecasting. Otherwise, the conclusions drawn from the analysis would be of little value.

To filter the data we follow and adapt the procedure by Genre et al. (2013). To this end, we split forecasts into *frequent* and *infrequent* respondents. The latter are than omitted from the panel. We label Forecasters infrequent respondents if they have more than four consecutive missing values, i.e., a year of non-responding. Because we use a rolling window, we have to process the data for every subset of observations separately. By that, experts that are labeled infrequent once can redeem themselves from past infrequent responding periods and will then contribute. Moreover, experts who have entered the SPF later in time have a chance to be considered and influence the combined forecast.

After filtering the panel, we now take a look at how to treat missing values. We previously ruled out the option to disregard all forecasts that have missing values. Alternatively, one can impute the missing values. This is a technique that tries to mimic forecasts behavior for the time period where no data is available and determine an appropriate value for it (see e.g., Efron, 1994). An intuitive requirement for data imputation is that there are sufficiently many actual observations. For example, if there is only one actual forecast out of a hundred required forecasts, it is not sensible to impute the 99 missing values. The data imputation procedure used in Genre et al. (2013) estimates the missing values by fitting an autoregressive model for each expert within the SPF and then forecast the next missing value. This, however, artificially changes both

[6] Note that we calculated this average response rates based on the number of missing submission after the first contribution of an expert.

the error variance of the experts and the forecast error covariances. However, the error covariance matrix is the crucial input for the forecast combination optimization problems and, as discussed in Section 2.3, estimation errors in the error variance covariance matrix contribute to the forecast combination puzzle. More importantly, our objective when analyzing the SPF is to gain insights into actual, real world error variances and covariances structures. If we impute forecast, this structure changes.

Therefore, we follow Matsypura et al. (2018) and Radchenko et al. (2023) and only use the actually observed data for the estimation of the error variance covariance matrix, i.e., estimate each element of the covariance matrix separately.[7] As a consequence, for the estimation of covariances we can only use points in time when responses of experts overlap. To ensure a minimal number of common observations, we additionally filter the panel once more disregarding all experts who responded less than 75% of the time for a given rolling window. By that, the number of common observations of two experts is in the worst case 50% of the window size, i.e., 22 observations which is six and a half years. We estimate the error variances based on at least 75% of the observations, i.e., 33 observations or eight and a quarter year.

Accordingly, we do not explicitly impute missing values or balance the panel. Instead, we only filter it and then use a different technique to estimate the error variance covariance matrix. The benefit of this approach is an error variance-covariance matrix that is only estimated based on actually observed data. The approach is also simply applicable to a real-world forecasting environment. For that, we extend the panel filtering and omit experts that have not contributed a forecast for the target period. Otherwise, we can determine a weight for these experts, but do not have an actual forecast for the target period that we can use to calculate the combined forecast.

3.1.3 Analysis of the Error Variance and Covariance of the SPF

The purpose of this analysis it to gather evidence and inspiration for the design of a simulation study. To this end, we analyze the error variances and error covariances / correlations of the SPF as those drive the solution to forecast combination problems. Figures 3.2 to 3.4 show both a heatmap and boxplots that are based on a subset of the rolling window approach of size 44. Within Figure 3.2 we use the first 44 observations of the data, i.e., from December 1999 to September 2010 which includes the financial crisis of 2007.

The heatmap in Figure 3.2 depicts the correlation matrix of the forecast errors.[8] Both the rows and columns correspond to experts with their forecaster ID, e.g., "F11". Each

[7] Note that this can result in covariance matrices that are not positive (semi) definite. Following Radchenko et al. (2023) we compute the nearest positive definite matrix using the *R* function *nearPD* from the *Matrix* package (D. Bates, Maechler, & Jagan, 2022; R Core Team, 2022).

[8] In order to calculate the forecast errors we downloaded the true inflation values from EuroStat `https://ec.europa.eu/eurostat/` on the 17th November 2021.

intersection is illustrated as a rectangle, and it represents the error correlation between the two experts. It is color-coded based on how high (or low) the error correlation is. Red colors correspond to higher error correlations while yellow indicate, in comparison, lower and green the lowest considered error correlations. The corresponding legend is on the rights side of the heatmap. The smallest error correlation (green) is about 0.69. The intermediate color yellows center is around a value of 0.82. More red color codes depict error correlations close to one. Due to the fact that the diagonal of an error correlation matrix is always one, it is grayed out such that it does not obstruct the analysis. Note that we fixed the color coding for all heatmaps within Figures 3.2 to 3.4. The ordering of the experts in the heatmap is based on their forecast error variance for the given subset of observations. The experts with the best forecast accuracy is at the top of the ordinate and the left side of the abscissa. Accordingly, the forecast with the largest MSE is on the bottom and right side respectively. Furthermore, we divide experts into four groups by the quartiles of the forecast accuracy. This is visualized by the four by four grids in Figure 3.2. For the sake of simplicity, we will name the experts with the smallest 25% of MSE values group one, the next group two and so on. The 25% with the largest MSE will be referred to as group four. On the top and right-hand-side of the heatmap there is an indicator to identify which visually differentiable rectangle belongs to which group. For example, the rectangle on the top left shows the error correlation between the 25% best forecast, i.e., group one, with experts from the same group. The rectangle on the bottom left (and top right) depicts the error correlation between group one and group four. In addition to the heatmap Figures 3.2 to 3.4 also include boxplots on the right side. They depict the MSE values (ordinate) for each group (abscissa). The rectangle of a boxplot depicts the middle 50% of observations, i.e., the values between the 75% and 25% quantiles. The black line within the rectangle is the median. The lines outside the rectangle are whiskers and either have the length of 1.5 times the length of the rectangle or up to the largest and smallest observation respectively. We will use these boxplots to analyze the error variance of the groups. Note that the boxplots are only based on six or seven observations.

Let us now take a closer look at Figure 3.2. Based on the heatmap one can see that for this set of observations the best experts are less correlated within their own group.[9] To an extent they are also less correlated with group two. In comparison, the error correlations between forecasts from group one and forecasts from group three as well as four are higher, i.e., more yellow- and red-dish. However, considering the second, third, and fourth groups reveals that the error correlations are higher *within*

[9] From heron forward we will use two linguistic simplifications. First, we will use the terms the best experts or worst forecaster. The best experts are those that belong to group one, i.e., the expert with the best forecast accuracy. Accordingly, for the worst forecaster or experts, i.e., those that have the largest error variance, are within group four. Second, note that unless indicated otherwise correlation refers to the error correlation and not the correlation between forecasts.

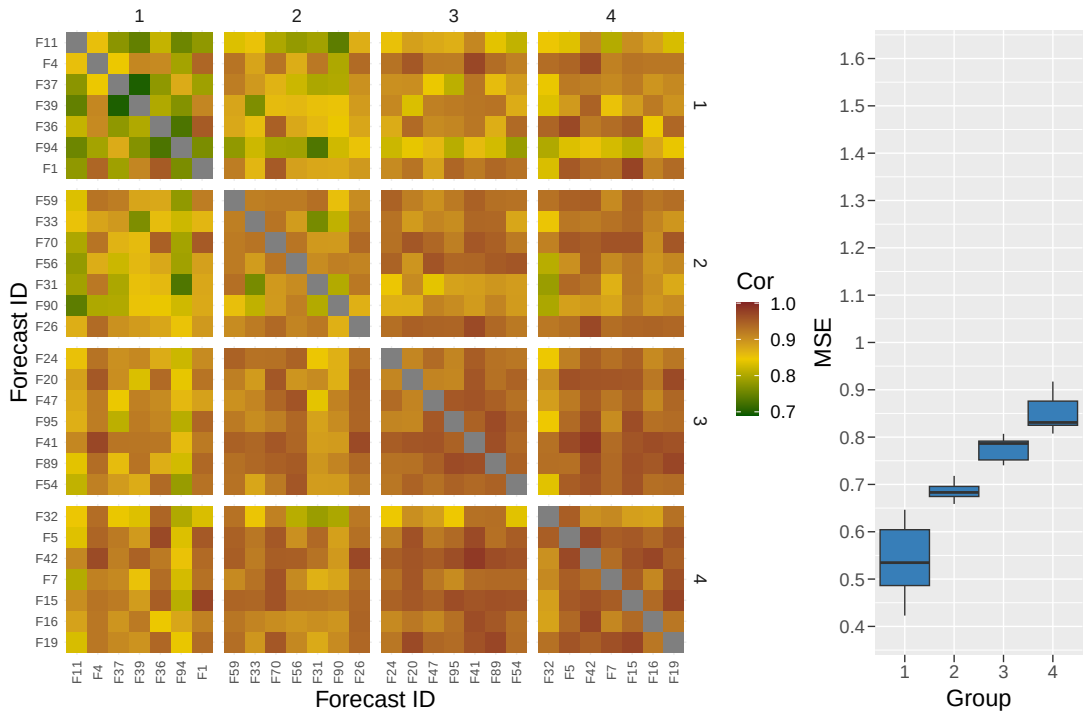


Figure 3.2. SPF forecast error variance and error correlation analysis from December 1999 to September 2010.

each group than *between* the corresponding group and the others. This is less prevalent for groups two to four, but clearly visible for group one. In general, one can observe that the error correlations without any differentiation between groups are high. The median error correlation is about 0.92 and only 7% of error correlations are below or equal to 0.80. As one can see due to the large amount of more green rectangles, a large proportion of the relatively smaller error correlations belong to forecasts from group one, particularly the error correlations between forecasts of group one. From the boxplot within Figure 3.2 we can see that group three and four have a similar median error variances (0.79 and 0.83). In comparison, the difference between groups one and two is 0.16 which is noticeably higher (median error variance 0.52 and 0.68). This hold similarly but to a lesser extent for group two and three with a difference of 0.11. Due to the fact that the boxplots are only based on a few values, we consider the whiskers to analyze the overall range of MSE values. One can see that the range of group one (0.38) is quite large in comparison to all other groups (0.06, 0.07 and 0.11).

In summary, group one is noticeably different from the other groups. First, forecasts from group one are less correlated with themselves as well as with other groups. Whereas for the other groups the correlations are higher both within and between groups. Second, the median error variance of group one is noticeably smaller compared to the other

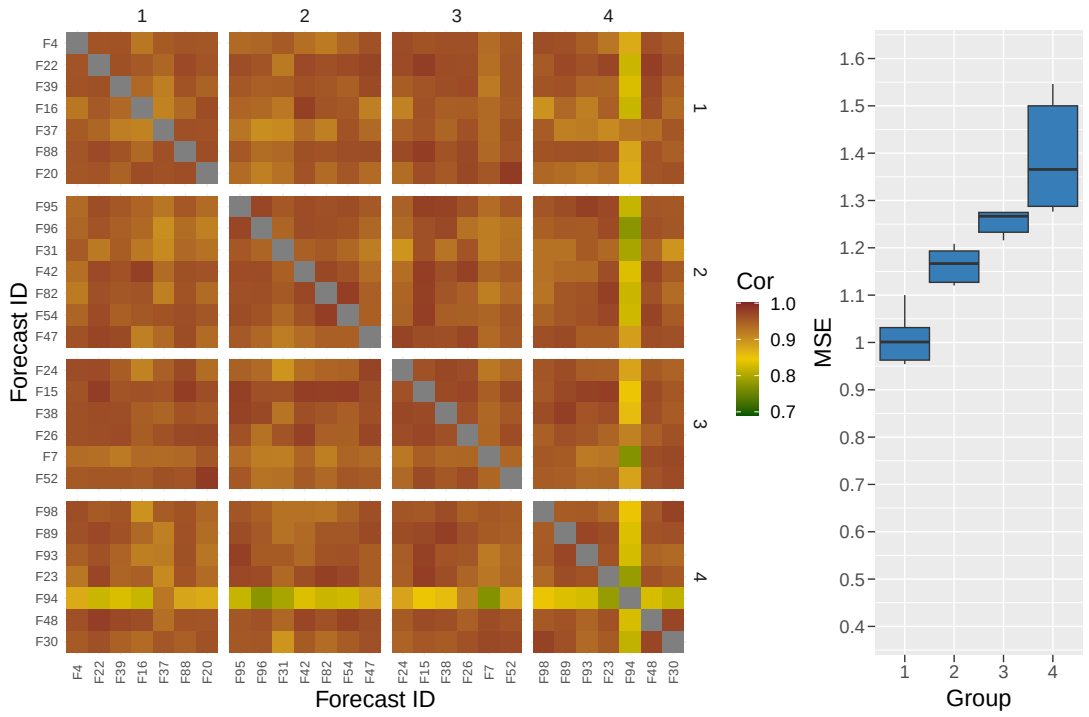


Figure 3.3. SPF forecast error variance and error correlation analysis from December 2005 to September 2016.

groups which median error variances are closer together. Lastly, the error variances within group one are more spread than all other groups.

Figure 3.3 depicts heatmap and boxplot for December 2005 to September 2016. For this subset of observations, all forecast errors are highly correlated. The median error correlation over all forecasts is 0.95 and over 92% of error correlations are greater 0.9. Based on the heatmap there are no significant differences between the error correlations among the four groups. There is, however, an anomaly clearly visible in group four. Forecaster "F94" is prominent as it is in comparison less correlated with any other expert. Interestingly, this behavior is observable throughout most subsets and even in Figure 3.2, although it is far less clearly noticeable. This expert showcases that there can be varying forecast accuracies throughout time. The expert starts in the first group and then transfers over group two and three into group four (Figure 3.3) as time goes by. Moreover, expert can vanish entirely from the subset of observation, see for example $F1$. Let us now consider the error variances of forecasts within Figure 3.3. Group one, two, and three are similar with regard to the size of their middle 50% and, to an extent, also with regard to their overall range. In contrast, group four differs from the first three as its range, middle 50% and overall error variance is noticeably larger. Similar to the observations of Figure 3.2, the difference in median error variance between groups is larger for group one (0.18) but this time also for group four (0.21). In comparison to

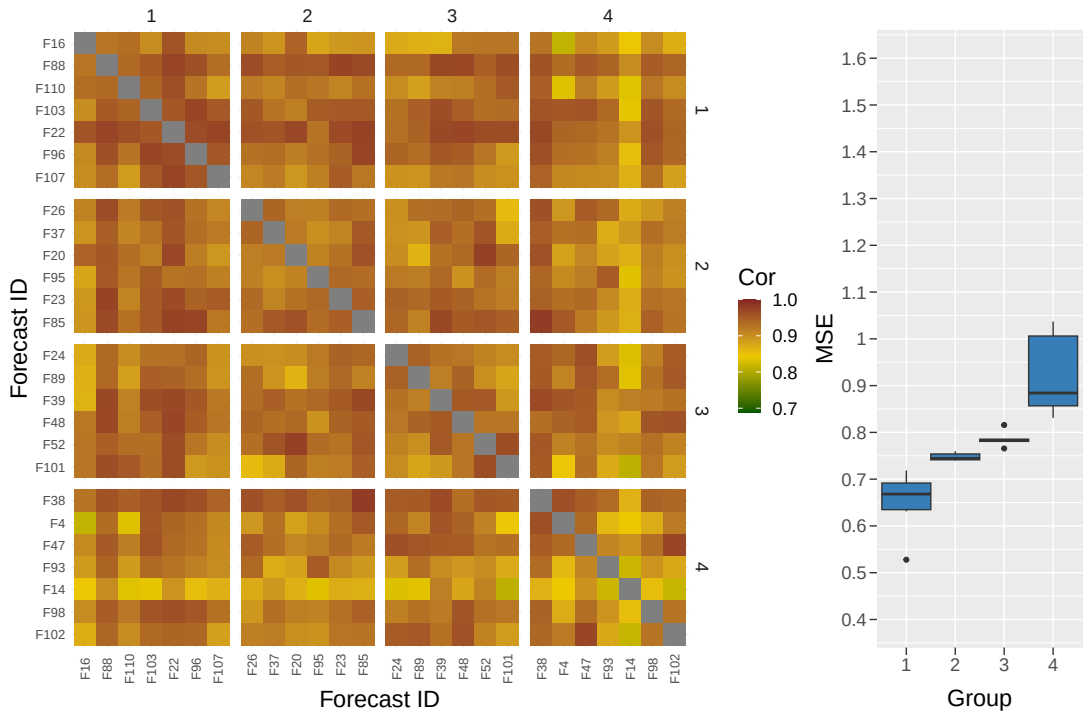


Figure 3.4. SPF forecast error variance and error correlation analysis from September 2010 to June 2021.

the first subset depicted in Figure 3.2, the overall error variance of the groups is higher, i.e., they have a worse forecast accuracy. In summary, we observe that the correlations are again all very high but this time also very similar across groups. Both the median error variances of group one and four show a larger difference two their neighboring groups median than groups two and three. Lastly, the range of error variances in the last group is noticeably larger.

For the last subset of observations we consider the time span from September 2010 to June 2021. Both a heatmap of error correlations and boxplots of forecast accuracy are depicted in Figure 3.4. The error correlations are in-between those shown in Figures 3.2 and 3.3. They are high overall, but there are also smaller error correlations present in comparison. However, those are still higher than what can be observed in Figure 3.2 for group one. Within the considered subset of observations the smaller error correlations are related to group four, both for experts within group four as well as between experts of group four and other experts from other groups. Taken the boxplot into consideration reveals that the overall forecast accuracy improved again, however, it is not at the same level as for the first considered subset. Interestingly, groups one and four both have a noticeably smaller and higher MSE respectively. Groups two and three have even closer median MSE values than before. However, groups two and three have very little variation in forecast accuracy as indicated by the thin rectangles, i.e., middle 50%.

Due to that, the whiskers are not long enough to cover all observations which are then depicted as dots, see group three. Group one also has such an outlier but, in this case, it is a noticeably better forecast.

In summary, throughout this analysis we found various structures for the error correlation and error variance. Overall, the forecast errors of different forecasts are highly correlated. The error correlations between groups can be very similar overall as in Figure 3.2. Alternatively, based on the presented data, the first or last group show some differences in form of smaller error correlations as depicted in Figures 3.2 and 3.3. With respect to the error variance or forecast accuracy of the forecast, there is evidence for varying differences between the median MSE of two neighboring groups, i.e., smaller and larger differences in the median MSE (Figure 3.4 versus Figures 3.2 and 3.3). Moreover, both the first and last group can noticeably differ, i.e., the difference in MSE is larger than between other groups. We use these results, among other things, to build a simulation study in Section 3.2. To this end, we will consider different groups of forecasts with varying forecast error variances and different error correlation structures.

3.2 Simulation Study

In order to analyze the forecast accuracy of different forecast combination methods we conduct a simulation study. To this end, we could either simulate the actual time series and simulate (or estimate) the corresponding forecasts (see e.g., Capistrán & Timmermann, 2009), or we could directly specify the variance-covariance matrix of the forecast errors, see Roccazzella et al. (2022). The former is preferable, if the impact of forecast model specification and estimation if of additional interest. The latter has the advantage, that it does neither rely on the assumption of particular forecasting models nor uncontrolled estimation error of the parameter values of the forecast models. Instead, it focuses directly on the specification of the most important input of the optimization problem: the variance-covariance matrix of forecast errors. Therefore, we use the second approach and extend the simulation setup of Roccazzella et al. (2022) to account for more sophisticated structures.

In Roccazzella et al. (2022) the variable of interest is given by

$$y \sim N(0, \sigma_y^2), \tag{3.1}$$

and the N unbiased forecasts of y are simply given by

$$\hat{y}_i = y + \varepsilon_i \qquad \forall\, i = 1, \ldots, N, \tag{3.2}$$

$$\text{with} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.3}$$

The $N \times 1$ vector $\mathbf{0}$ contains only zeros and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of forecast error terms ε_i. The error terms are drawn from a multivariate normal distribution with zero means and an error variance-covariance matrix $\boldsymbol{\Sigma}$. Thereby, the most important part within this simulation study is the error variance covariance matrix $\boldsymbol{\Sigma}$. It defines both the error variance (or forecast accuracy) for each forecast and the error correlation between them. In Roccazzella et al. (2022) forecasts are divided into two groups: $\mathbb{G}_1$ and $\mathbb{G}_2$. Within each group the error variance of the forecasts are identical. Forecasts in $\mathbb{G}_1$ have an error standard deviation that is 50% smaller compared to forecasts in $\mathbb{G}_2$.[10] For the error correlation Roccazzella et al. (2022) chose specific value between 0.01 up to 0.9. The error variances and covariances are computed as follows:

$$\sigma_i^2 = \begin{cases} \alpha^2 & i \in \mathbb{G}_1 \\ (2\alpha)^2 & i \in \mathbb{G}_2 \end{cases} \tag{3.4}$$

$$\sigma_{i,j \neq i} = \rho_{i,j} \sqrt{\sigma_i^2 \sigma_j^2} \tag{3.5}$$

We adopt and extend this framework for our simulation study. Firstly, we extend the design of Roccazzella et al. (2022) by introducing S different groups of forecasts with $S \geq 1$. Each group is denoted by $\mathbb{G}_s$ with $s = 1, \ldots, S$. The number of total forecasts is given by N, and it has to hold that $N \geq S$. For now, we assume that the number of forecasts within each group N/S is identical and an integer.

Beside the number of groups S and the number of forecasts N, there are basically two aspects that shape each of our scenario. First, the specification of the variance and, second, the specification correlation of the forecast errors. The error covariances are then calculated by Equation (3.5). In contrast to Roccazzella et al. (2022) we determine the error variance and correlation differently. We will discuss this in Section 3.2.1 and present a general, more comprehensive framework to create simulation studies to analyze forecast combination methods. We will present various ways in which the framework can be changed to allow for the design of extensive and diverse simulation studies. For example, it allows for more diversified groups, i.e., not all forecasts of the same group share the same error variance and correlation. Thereafter, we will present the specific scenarios that we will consider throughout this thesis in Section 3.2.2. Lastly, Section 3.2.3 summarizes the simulation framework and the scenarios that will be considered.

[10] In another scenario forecasts from $\mathbb{G}_2$ are random noise.

3.2.1 General Simulation Framework for Analyzing Forecast Combination Methods

First, we want to create the forecast error variance. To this end, we sort forecasts based on the error variance increasingly not only between groups but also within groups. Thus, group one ($s = 1$) consists of forecasts with the lowest error variances whereas group S contains the forecasts with the highest error variances. We define $N/S \times 1$ dimensional vectors $\boldsymbol{\sigma}_s^2$ for all $s = 1, \ldots, S$. Each vector $\boldsymbol{\sigma}_s^{\mathbf{2}}$ contains the error variance σ_i^2 of each forecast $i \in \mathbb{G}_s$. It holds that

$$diag(\boldsymbol{\Sigma}) = (\boldsymbol{\sigma}_1^{2\prime}, \boldsymbol{\sigma}_2^{2\prime}, \ldots, \boldsymbol{\sigma}_S^{2\prime}). \tag{3.6}$$

To determine the error variance $\boldsymbol{\sigma}_s^{\mathbf{2}}$ of each group, we first define the *median error variance* of a group and then, based on this, calculate the *forecasts error variances* in a group. By that, there are two adjustments that can be made. First, one can specify how groups from a *global perspective* perform in comparison to each other. Second, to consider the performance of forecasts from a *local perspective*, we can alter the distribution of error variances within groups without changing the global perspective with regard to their median error variances.

Median Error Variance First, consider the global perspective. The *median error variances* of each group, $\eta_s \ \forall \ s = 1 \ldots, S$, are contained in the $S \times 1$ vector $\boldsymbol{\eta}$ which is determined by

$$\boldsymbol{\eta} = \eta_1 \boldsymbol{\alpha}. \tag{3.7}$$

In general, the base median error variance η_1 can be used to define different magnitudes of error variances. However, without loss of generality we set $\eta_1 = 1$ in our simulation study, i.e., the median error variance of group one is identical to the base median error variance. Thereby (3.7) simplifies to $\boldsymbol{\eta} = \boldsymbol{\alpha}$. Thus, the $S \times 1$ vector $\boldsymbol{\alpha}$ represents not only the relative difference between the median error variance of any group to the first group but also the median error variances themselves. It is given by

$$\boldsymbol{\alpha}' = (1, 1 + z_2, 1 + z_3, \ldots, 1 + z_S), \tag{3.8}$$

with $z_s < z_{s+1} \ \forall \ s = 2, \ldots, S-1$ and $z_s > 0 \ \forall \ s = 2, \ldots, S$.[11] Because $\eta_1 = 1$, it follows that $z_s \cdot 100\%$ is the relative distance of the median error variance of group s compared to group one. In what follows we will refer to z_s as the *error variance similarity*. By defining different values for $z_s \forall \ s = 2, \ldots, S$ one can introduce a variety of scenarios.

[11] We chose to start indexing at two for the sake of clarity regarding which z-value belongs to which group and because z_1 is always zero. The median error variance of group one is defined by η_1^{med} in equation (3.7).

For example, if z_2 is significantly larger than $z_3 - z_2$ one creates a scenario where the best group (group one) provides noticeably better forecasts while the groups two and three have a more similar forecasting performance.

Group Error Variances Second, we consider the local perspective. Henceforth, we will refer to the error variances within the groups by *Group Error Variances*. Based on the median error variances, $\boldsymbol{\eta}$, we determine the error variances of forecasts within a group according to

$$\boldsymbol{\sigma}_s^2 = \eta_s + \boldsymbol{\beta}_s \qquad s = 1, \dots, S, \tag{3.9}$$

with η_s being the s-th vector element of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}_s \in \mathbb{R}^{|\mathbb{G}_s|}$. The vector $\boldsymbol{\beta}_s$ determines the error variances of each forecast in group s relative to the median error variance η_s. If the $k - th$ element of $\boldsymbol{\beta}^s$, i.e., $\beta_{s,k}$, is smaller than zero, this implies the forecast k from group s has a smaller error variance than the median error variance η_s. For values greater zero the error variance is higher compared to η_s.

The vector $\boldsymbol{\beta}^s$ is constraint to some extent. First, because we assumed that forecasts are sorted based on the error variance within each group non-decreasing, the elements in $\boldsymbol{\beta}^s$ have to be non-decreasing. Second, we define that all forecasts within group s have a smaller error variance than forecast within group $s + 1$. Accordingly, it has to hold that

$$max(\boldsymbol{\beta}_s) + \eta_s < min(\boldsymbol{\beta}_{s+1}) + \eta_{s+1} \qquad \forall\ s = 1, \dots, S - 1. \tag{3.10}$$

Third, if the number of forecasts in a group, $|\mathbb{G}_s|$, is odd the median error variance η_s itself is by definition the middle element $\boldsymbol{\sigma}_s^2$. To be more precise, it will be the k^*-th entry of $\boldsymbol{\sigma}_s^2$ with $k^* = \lceil |\mathbb{G}_s|/2 \rceil$. Therefore, the $k^* - th$ element of $\boldsymbol{\beta}^s$ has to be set to zero. If, instead, the number of forecasts in a group is even, the median is the average of the k^*-th and $(k^* + 1)$-th element of $\boldsymbol{\sigma}_s^2$. To ensure that the median error variance η_s is unchanged the k^*-th and $(k^* + 1)$-th element of $\boldsymbol{\beta}^s$ can not be chosen separately. In summary, it has to hold that

$$\beta_{s,\lceil k^* \rceil} = \begin{cases} 0 & \text{if } k^* = |\mathbb{G}_s|/2 \text{ is odd} \\ -\left(\beta_{s,(k^*+1)}\right) & \text{if } k^* = |\mathbb{G}_s|/2 \text{ is even} \end{cases} \qquad \forall\ s = 1, \dots, S. \tag{3.11}$$

Subject to the above constraints, $\boldsymbol{\beta}_s$, can be determined both by hand or more automatically.

In order to determine $\boldsymbol{\beta}^s$ automatically, we need to impose (temporary) conditions.[12] First, let the elements of $\boldsymbol{\sigma}_s^2$ $\forall s = 1, \ldots, S$ be equidistant. Second, let there be a constant increase in the median error variance from any two neighboring groups, i.e., $z_s = (s-1)z \; \forall \; s = 2, \ldots, S$ with $z = z_2$. As a result, the vector $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}' = (1, 1+z, 1+2z, \ldots, 1+(S-1)z). \tag{3.12}$$

For the second condition, the difference between $max(\boldsymbol{\sigma}_s^2)$ and $min(\boldsymbol{\sigma}_{s+1}^2)$ has to be the same as the difference between any two neighboring elements within $\boldsymbol{\sigma}_s^2$ for any $s = 1, \ldots, S$. As a consequence $\boldsymbol{\beta}^s$ is identical for all $s = 1, \ldots, S$ and the k-th element of $\boldsymbol{\beta}^s$ is given by

$$\beta_{s,k} = \frac{zS\Big(k - \Big\lceil \frac{N}{2S} \Big\rceil - 0.5\Big(1 - \frac{N}{S} \bmod 2\Big)\Big)}{N}. \tag{3.13}$$

The vector $\boldsymbol{\beta}^s$ can be altered further to create specific scenarios. For example, extraordinary good or particularly bad forecasts can be introduced, i.e., outliers in the first and / or last group.

With the median error variances contained in $\boldsymbol{\eta}$ and the vectors $\boldsymbol{\beta}_s$, we can determine all error variances, i.e., the diagonal of the error variance covariance matrix, by Equation (3.9).

Correlation In order to determine the error covariances, i.e., the off-diagonal elements of the error covariance matrix, we use Equation (3.5). To this end, we consider the error correlation matrix. The error correlation between any two forecasts $i, j = 1, \ldots, N$ of any two groups $s, r = 1, \ldots, S$ is denoted by $\rho_{i,j}^{s,r}$. Again, one can set the $\frac{N(N+1)}{2} - N$ error correlations to specific values of interest. Alternatively, one can define an error correlation of group r with each other group s as $\bar{\rho}^{r,s}$ $\forall \; r, s = 1, \ldots, S$. Either one can and use it for the whole group, i.e.,

$$\rho_{i,j}^{s,r} = \bar{\rho}^{r,s} \quad \forall \; r, s = 1, \ldots, S, \quad \forall \; i \in \mathbb{G}_s, \; j \in \mathbb{G}_r, \tag{3.14}$$

or an interval around $\bar{\rho}^{r,s}$ can be defined. Then the specific values of $\rho_{i,j}^{s,r}$ can be randomly drawn from the interval, or they can be assigned a value from an equidistant grid ranges within the bounds of the interval. Furthermore, the correlations can be hand-picked or drawn randomly from a given distribution or interval.

[12] Note that those conditions are needed to create $\boldsymbol{\beta}^s$ but afterwards they may not hold if, for example, $\boldsymbol{\beta}^s$ is further changed by hand or if the median error variance of the first or last group is changed.

3.2.2 Designing Scenarios for the Simulation Study

In this section we will further discuss what scenarios will be designed using the simulation study presented. Although, we provide a framework for a simulation study that can create a variety of different scenarios for forecast combination, we have to narrow down the scenarios we can consider. Based on this general framework of the simulation study, we will now present the specific scenarios that we consider in this thesis.

For the design of the simulation study we, henceforth, use $\eta_1 = 1$, i.e., (3.7) simplifies to $\boldsymbol{\eta} = \boldsymbol{\alpha}$. Additionally, we consider $N = 24$ forecasts, $S = 4$ groups and the variance of the true series is $\sigma_y = 1$. By that we have a large pool of potential forecasts for the forecast combination methods. The larger number of forecasts has multiple reasons. First, in conjunction with a short time series, it increases the estimation uncertainty. Recall, that in Section 2.3 we discussed that the estimation uncertainty / error is a potential reason why sophisticated forecast combination method are outperformed by simple combination methods. Thus, we can examine how forecast combination methods perform in environments of larger uncertainty. Second, a larger number of forecasts is also used in other studies regarding forecast combination (see e.g., Capistrán & Timmermann, 2009; Roccazzella et al., 2022). Third, the SPF has a larger number of forecasts, and we took inspiration from that. Finally, this provides a pool of forecasts to chose from for methods that also perform a variable selection.

In Section 3.2.2.1 we will define which scenarios for the error variance, σ_i^2, we will take into consideration. Thereafter, we will consider the correlations between the forecast errors in Section 3.2.2.2.

3.2.2.1 Error Variance

In order to specify the error variance of forecasts, we need to determine the *group error variances* $\boldsymbol{\beta}^s \; \forall \; s = 1, \ldots, S$ and the *median error variance* $\boldsymbol{\alpha}$ and η_1, see Equations (3.7) and (3.9).

Group Error Variance For the group error variance defined by $\boldsymbol{\beta}^s$ we will use the automated approach described in Section 3.2.1, i.e., $\boldsymbol{\beta}^s$ is identical for all $s = 1, \ldots, S$ and each element β_k^s is determined by Equation (3.13).

Median Error Variance Recall, that for the automated approach to define $\boldsymbol{\beta}^s$, we imposed conditions. As a result, we have to use $z_s = (s-1)z \; \forall \; s = 2, \ldots, S$ with $z = z_2$. It means that there is a constant increase in error variance compared to group one. Each group s has a $(s-1)z \cdot 100\%$ larger median error variance η_s compared to η_1 $\forall \; s = 1, \ldots, S-1$. Accordingly, for $S = 4$, the vector $\boldsymbol{\eta} = \boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha} = (1, 1+z, 1+2z, 1+3z)'. \tag{3.15}$$

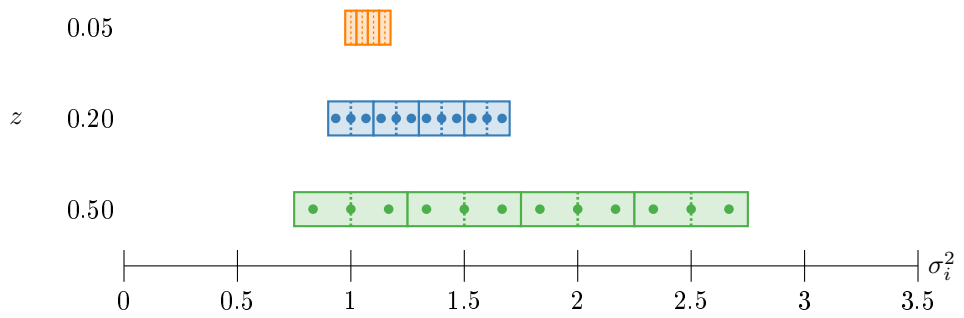


Figure 3.5. Illustration of median error variances and group error variances with respect to the error variance similarity (z) and no special groups.

For the error variance similarity we use $z \in \{0.05, 0.2, 0.5\}$. Similar values have, for example, been used in Blanc and Setzer (2020) for individual forecast error variances. Furthermore, we observed more similar as well as more dissimilar forecast accuracies when analyzing the SPF within Section 3.1.3. However, we have to point out that, although we build upon those insights, we simplify the simulation study by not using individual values $z_s \; \forall \; s = 2, \ldots, S$ but constant ones. Figure 3.5 illustrates the error variances of both the groups and forecasts depending on the choices we made for the group and median error variance. The abscissa shows the error variances and the ordinate represents the three error variance similarities, z-values, that are used. The coloring is used as an additional aid for the different values of z and will become more relevant later on. The four different groups of forecasts are depicted by the rectangles. Within each groups middle there is a dashed line that represents the corresponding median η_s. Since we chose $\eta = 1$, the dashed line of the first group is always at $\sigma^2 = 1$. The dots within each rectangle represent the individual error variances of the forecasts. For this example we used $N = 12$ forecasts, i.e., three forecasts per group. By choosing $z \in \{0.05, 0.2, 0.5\}$ we create a variety of scenarios. With $z = 0.05$ (orange) we create a scenario where both the groups are close in terms of their median error variance, but also the forecasts error variances are very similar. Note that for $z = 0.05$ the individual forecast error variances are not shown because of the narrowness of the rectangles. As we increase z to 0.2 (blue) and 0.5 (green) we create groups that are less similar or more dissimilar with regard to their median error variances (the dashed lines). By that, the difference between two forecasts error variances (dots within the rectangles) also increases. Overall, the median error variances for $z = 0.05$ are between 1 and 1.15, i.e a 15% difference between the median error variance of the best and worst group. For $z = 0.2$ the relative difference is 60%, i.e., the median of group four is 2.5 times larger than that from group one. For $z = 0.5$ we have a 150% difference or η_4 is 2.5 times larger than η_1 or respectively.

Figure 3.5 also illustrates how the forecasts are distributed within each group. This distribution is the result from the automated approach to determine $\boldsymbol{\beta}^s$. We see that

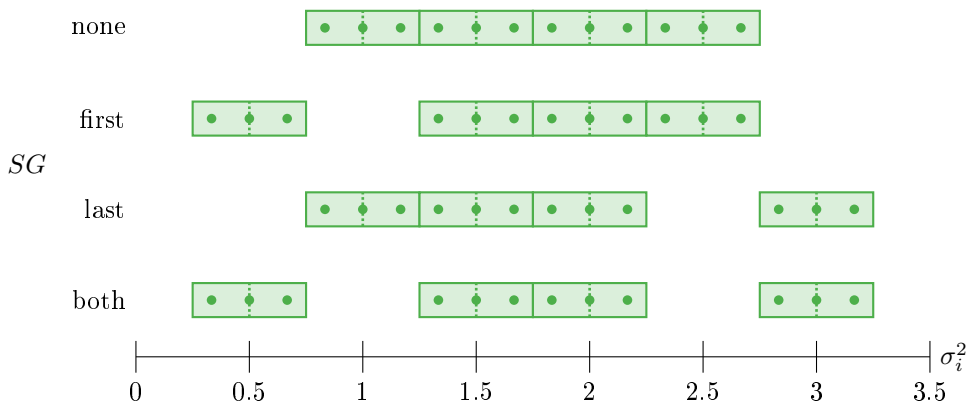


Figure 3.6. Illustration of median error variances and group error variances with respect to special groups (SG) and $z = 0.5$.

they are equidistant within each group. Additionally, two neighboring forecasts from two different groups have the same distance. Recall that the difference between the forecast error variances also increases when z increases. By that, the groups are more diverse or dissimilar as z increases.

We presented how we define the error variances σ_i^2 for each forecast $i = 1, \ldots, N$ based on the median error variance, Equation (3.15), and group error variance, Equation (3.13). By that, we created three different scenarios: similar ($z = 0.05$), less similar ($z = 0.2$) and dissimilar ($z = 0.5$) groups and forecast error variances. On top of that we want to introduce *special groups (SG)*. The idea of special groups is based upon the insights gained in section 3.1.3, more precisely based on Figures 3.2 to 3.4. We consider scenarios in which whole groups perform significantly differently compared to other neighboring groups. To this end, we consider three different variations of the median error variance, i.e., variations of $\boldsymbol{\eta} = \boldsymbol{\alpha}$, see Equation (3.12). For the sake of simplicity, we will differentiate the scenarios by the considered "special groups". Figure 3.6 illustrates the four different scenarios for special groups. Similar to Figure 3.5 the abscissa still represent the error variances. The ordinate shows the four scenarios of special groups: *none*, *first*, *last*, *both*. For the sake of better visibility, we use $z = 0.5$, i.e., we showcase the special groups for the green colored error variances or rectangles of Figure 3.5. The first case of SG *none* is the base scenario we already considered. No group has a significantly better or worse forecast accuracy. To put it differently, the median error variances of neighboring groups are within the same distance from each other. If SG *first*, the best group of forecasts, is noticeably set off from all other groups, i.e., has an even better forecast accuracy. By that, there is a gap between the error variances of group one and two as illustrated in Figure 3.6. If SG is *first*, we reduce the base median error variance η_1 by z. Accordingly, the difference in the median error

variance between group one and two, is $2z$. The corresponding vector $\boldsymbol{\alpha}$ to determine the median error variances $\boldsymbol{\eta}$ is given by

$$\boldsymbol{\alpha}' = \big(1 - z, 1 + z, 1 + 2z, 1 + 3z\big). \tag{3.16}$$

Note that the distribution of forecasts within the groups remains unchanged, see Figure 3.6.[13] The MSE values of forecasts within groups are still equidistant, but the difference between the forecasts with the largest error variance of group one and the forecast with the smallest error variance of group two differs.

Next we consider the scenario, SG *last* where the last group of forecasts with the largest median error variance gets even worse. By that, we create a gap between the last and all other groups, see Figure 3.6. To this end, we define

$$\boldsymbol{\alpha}' = \big(1 + z, 1 + z, 1 + 2z, 1 + 3z + z\big). \tag{3.17}$$

Lastly, the scenario SG *both* combines SG *first* and SG *last*. The best forecasts are even better and the worst forecasts are even worse. We define

$$\boldsymbol{\alpha}' = \big(1 - z, 1 + z, 1 + 2z, 1 + 3z + z\big). \tag{3.18}$$

In summary, with $z \in \{0.05, 0.2, 0.5\}$ we define scenarios of similar, less similar and dissimilar forecasts. With SG *none*, *first*, *last*, *both* we further introduce scenarios within groups of forecasts that are significantly better or worse in comparison to the other groups. We believe that those scenarios are relevant and interesting, and it is worth analyzing how the forecast combination methods perform within them. However, if we use SG *first*, there is a group that is noticeably better and more importantly, we introduce forecasts with a lower error variance, see again Figure 3.6. Therefore, for a comparison between scenarios with SG *none* and SG *first*, we expect that, overall, the MSE of all methods is reduced. This has to be taken into consideration when comparing scenarios.

Within this section we created multiple scenarios for the error variances of forecasts. Next we need to specify which error correlations we want to use.

3.2.2.2 Correlation Matrix

The scenarios for the error variances of forecasts that we defined in Section 3.2.2.1, are the diagonal elements of the error variance covariance matrix $\boldsymbol{\Sigma}$. In order to determine the off-diagonal elements, i.e., error covariances, we first define the error correlations. Then the error covariances can be calculated by Equation (3.5).

[13] Further, note that the third condition for the automated approach for $\boldsymbol{\beta}^s$ is needed to create beta. This, however, does not imply that these assumptions have to hold after further manipulation of $\boldsymbol{\alpha}$.

Within Section 3.2.1 we presented different possibilities to determine correlation matrices. In terms of our simulation study we chose to use fixed correlation matrices, i.e., we neither draw the common error correlation randomly nor do we determine the individual error correlations randomly based on $\bar{\rho}^{r,s}$. We have decided against this for reasons of reproducibility of our results. Furthermore, hand-picking all correlations separately creates very specific error correlation matrices. When considering the heatmaps of the SPF within Section 3.1.3, we discussed that the error covariances of a group are oftentimes similar, however not identical. We build upon these insights but simplify it for our simulation study. Henceforth, we will use a fixed correlation for all error correlations. Note that this is already an extension to Roccazzella et al. (2022) where the same correlation was used for all forecast errors. Accordingly, the error correlation of forecast i from group r with forecast j from group s with $r, s \in 1, \ldots, 4$ is given by

$$\rho_{i,j} = \bar{\rho}^{r,s} \qquad \forall i, j \in \{1, \ldots, N : i \neq j\}. \tag{3.19}$$

Because we define the individual error correlation based on $\bar{\rho}^{r,s}$, we only need to define a 4×4 (error) correlation matrix (CM). Each entry represents the error correlation for the whole group, i.e., $\bar{\rho}^{r,s}$ $\forall$ $r, s \in 1, \ldots, 4$.

Our analysis based on the SPF as well as similar findings in the literature suggest that highly correlated forecast errors are common. It can be a result of forecast being based on similar methods and data sets (see e.g Wang et al., 2023; Winkler & Clemen, 1992). Nevertheless, we want to create a diverse set of scenarios. Therefore, the first three correlation matrices represent scenarios of *high*, *medium*, and *low* (positive) error correlations. To this end, we define an error correlation coefficient of 0.9 as high, 0.5 as medium and 0.2 as low. Accordingly, the first three correlation matrices are

$$CM1 = \begin{pmatrix} 0.90 & 0.90 & 0.90 & 0.90 \\ & 0.90 & 0.90 & 0.90 \\ & & 0.90 & 0.90 \\ & & & 0.90 \end{pmatrix} \quad CM2 = \begin{pmatrix} 0.50 & 0.50 & 0.50 & 0.50 \\ & 0.50 & 0.50 & 0.50 \\ & & 0.50 & 0.50 \\ & & & 0.50 \end{pmatrix}$$

$$CM3 = \begin{pmatrix} 0.20 & 0.20 & 0.20 & 0.20 \\ & 0.20 & 0.20 & 0.20 \\ & & 0.20 & 0.20 \\ & & & 0.20 \end{pmatrix} \tag{3.20}$$

In addition to that we will use three more correlation matrices that are, partly, inspired by our analysis of the SPF. They are given by

$$CM4 = \begin{pmatrix} 0.90 & 0.50 & 0.50 & 0.50 \\ & 0.90 & 0.50 & 0.50 \\ & & 0.90 & 0.50 \\ & & & 0.90 \end{pmatrix} \quad CM5 = \begin{pmatrix} 0.90 & 0.50 & 0.50 & 0.20 \\ & 0.50 & 0.50 & 0.20 \\ & & 0.50 & 0.20 \\ & & & 0.20 \end{pmatrix}$$

$$CM6 = \begin{pmatrix} 0.50 & 0.50 & 0.50 & 0.50 \\ & 0.90 & 0.90 & 0.90 \\ & & 0.90 & 0.90 \\ & & & 0.90 \end{pmatrix} \tag{3.21}$$

The fourth correlation matrix $CM4$ is not based upon observations form the SPF. Instead, we wanted to design a scenario where forecasts from the same group are highly correlated, i.e., more similar, but are less related to forecasts from other groups. For $CM5$ we took inspiration from the error correlations presented in Figure 3.4 but exaggerated it. Forecast errors from within the last, worst group are less correlated with themselves and forecast errors from other groups (low correlation). To put it differently, they, in contrast to the other groups, either use different information or extract the information differently. The second and third group are more similar both within the groups and between them (medium correlation), i.e., they use more important and similar information and extract the information similarly efficient. The best group has very similar forecast errors, i.e., high correlation. We interpret this as if all of them use a large portion of the same important information and extract it similarly to accurate forecast. In contrast, $CM6$ considers a contrary scenario. It is inspired by the error correlations of forecasts of the SPF shown in Figure 3.2. Groups two, three, and four have a similar high correlation. They use similar information but their effectiveness in extracting it differs which drives the different forecast accuracies. Group one has the best forecasts. However, the error correlation among forecasts within group one as well as with forecasts from other groups is medium. It is a scenario where each forecast in group one found a different aspect that elevates their forecast accuracy or extracted important parts of the common information more effectively.

3.2.3 Brief Summary of the Designed Scenarios

In this chapter, we extend the simulation study of Roccazzella et al. (2022) and, thereby, present a framework for simulations studies to analyze the performance of forecast combination methods which is the first main contribution of this thesis. It allows multiple groups of forecasts which can have different magnitudes of error variances and

distributions of error variances. Furthermore, it allows to define different correlations between the forecast errors of different groups. To this end, one needs to define the error covariance matrix directly which is the input into a forecast combination problem. Section 3.2.1 presented the framework that can be used as well as several configurations or approaches to determine the median error variance, group error variance and error correlations.

Inspired by, inter alia, the analysis the ECB's Survey of Professional Forecasters we specified different concrete scenarios for our simulation study. All of them use $N = 24$ forecast and $S = 4$ groups. The median error variances as well as error variances within a group range from similar over less similar to dissimilar ($z \in \{0.05, 0.2, 0.5\}$). We can introduce special groups, i.e., the best forecast has an even better forecast accuracy, the last, fourth groups forecast accuracy is even worse or both (SG *none*, *first*, *last*, *both*). Lastly, we use high, medium, and low error correlations as well as some more specific design that are build around the idea of using different information or varying effectiveness in processing information.

4 L_1 Norm Constraints

To showcase that additional constraints are beneficial for forecast combination, we consider an analysis by Clemen (1986). In Chapter 2, we showed that forecast combination can also be represented as a regression model which was suggested by Granger and Ramanathan (1984). It is an approximation of the forecast combination problem by J. M. Bates and Granger (1969). Recall that an unconstrained regression model, as depicted in Equation (2.27), minimizes the in-sample mean squared error (MSE). In other words, there are no other weights that result in a smaller MSE. Clemen (1986) demonstrated that if we assume unbiased forecasts, we can improve the out-of-sample MSE if we use constraints. To this end, they used the regression model considered by Granger and Ramanathan (1984) which constrains the intercept to be zero and the weights to sum to unity.[14] Although, the unconstrained regression model minimizes the in-sample MSE, if forecasts are unbiased the out-of-sample MSE of the constrained regression model is smaller than that of the unconstrained regression model. However, if the forecasts have a small bias, the combined forecast based on the constrained regression is itself biased. Nevertheless, even in such cases, the constrained regression can still have a superior out-of-sample MSE if the increase in bias is sufficiently small (Clemen, 1986).

In this chapter, we introduce L_1 norm constraints and analyze its potential to improve forecast accuracy. To this end, we use it as an additional constraint that we impose onto the forecast combination problem introduced by J. M. Bates and Granger (1969) depicted in Equation (2.22). Recall, that the objective function of the forecast combination problem is to minimize the error variance of the combined forecast under the unity constraint, i.e., all weights sum to one.[15] In this chapter we analyze how an additional L_1 norm constraint can be imposed in various forms to reduce the out-of-sample MSE of the combined forecast. Note that the L_1 methods for forecast combination are closely related to the well-known *Least Absolute Shrinkage and Selection Operator (Lasso)* introduced in Tibshirani (1996). Therefore, we will also refer to these methods as lasso-based forecast combination methods.

[14] It should be noted, that in this case the MSE is approximately equal to the error variance if forecasts are unbiased.

[15] Note while in a regression framework like analyzed in Clemen (1986) one can consider forecast combination without any constraint this is not sensible for the combined forecast error variance minimization approach from J. M. Bates and Granger (1969). In the absence of the unity constraint, the objective function (error variance of the combined forecast) is minimized when zero weight is assigned to all forecasts.

The objectives for this chapter are the following:

(I) We introduce shrinkage and demonstrate how it can improve the out-of-sample forecast accuracy.

(II) We present various implementations and variations of the L_1 constraint that are used in the literature and analyze their impact on the weight vector.

(III) We develop a forecast combination problem that encompasses all the considered L_1 constraints.

(IV) We propose to use *Conditional Group Equal Weights (CGEW)* as a shrinkage direction for forecast combination with L_1 constraints.

(V) We develop a forecast combination problem that encompasses all the considered L_1 constraints.

(VI) We analyze the performance of the methods within an extensive simulation study for various scenarios, both with and without hyperparameter estimation.

With respect to the overall structure of this thesis, this chapter includes the second and third major contributions stated in Chapter 1: a unified framework for the L_1 constraint methods and Conditional Group Equal Weights as a shrinkage direction.

The remainder of this chapter is organized as follows. First, in Section 4.1 we introduce the concept of shrinkage, motivate its general idea and show how it can reduce the forecast accuracy of the combined forecast. We also introduce a related shrinkage method called Linear Hybrid Shrinkage (LHS) to compare its forecast accuracy to the L_1 constraint or lasso-based methods. However, the L_1 methods are our main focus. Furthermore, this section presents different variants and implementations of the L_1 constraint as they are considered in the literature.

In Section 4.2 we translate all variants and implementations of the L_1 constraint into a unified framework that is based on the original forecast combination problem proposed by J. M. Bates and Granger (1969). This framework minimizes a quadratic function, the error variance of the combined forecast, subject to linear constraints. As a result the variants of the L_1 constraint can be compared on the same basis. Furthermore, we analyze how each L_1 method affects the weight vector and present an optimization problem that nests all considered variants of the L_1 constraint. This includes shrinkage towards a prior weight vector.

In Section 4.3 we analyze the forecast combination methods within an extensive simulation study. Within this simulation study we analyze the performance both with and without the uncertainty of estimating a hyperparameter which is required for the L_1 methods.

Lastly, in Section 4.4 we discuss our results and directions for future research.

4.1 Background and Related Literature

In this section we motivate the idea of shrinkage and show its benefits in Section 4.1.1. In Section 4.1.2 we present L_1 variants and their implementation that are used for forecast combination in the related literature.

4.1.1 The Idea and Benefits behind Shrinkage

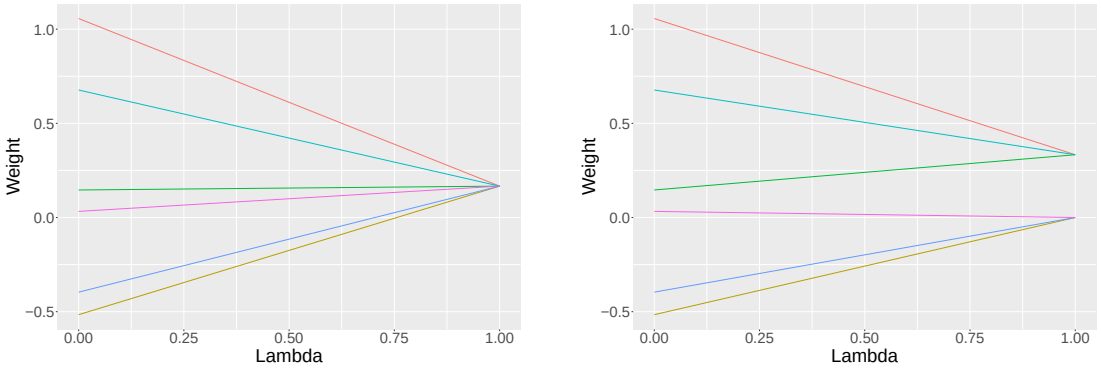
This thesis is centered around the concept of shrinkage (or regularization). Shrinkage is a technique that alters estimates towards a certain reference and by that can reduce the estimation error or variance (G. James et al., 2023; W. James & Stein, 1992). It has been used to reduce the squared error of the mean vector estimation for multivariate normal distributions (W. James & Stein, 1992; Stein, 1956; Tsukuma & Kubokawa, 2020, particularly pp. 2-5). Furthermore, it is used in the context of covariance estimation, for example in portfolio selection where the number of variables or assets can be very large (see e.g., Ledoit & Wolf, 2017). It is also used for machine learning or deep learning to avoid overfitting the training set (Aggarwal, 2023, pp.178-182). We consider it in the context of forecast combination with constrained weights.

Consider a commonly known multivariate regression problem in the form of Equation (2.27). The ordinary least squares estimator minimizes the sum of in-sample squared errors for a given data set. In fact, the ordinary least squares (OLS) estimator has the smallest in-sample error variance (or MSE) of all unbiased linear estimators. This is known as the Gauss-Markov theorem (see e.g., Hastie, Tibshirani, & Friedman, 2009, pp.51-52). However, the OLS estimator does not necessarily produce the best forecast $\hat{y}$. The out-of-sample error variance or MSE can be decomposed as follows:

$$E[(y-\hat{y})^2] = \underbrace{E[\hat{y}-y]^2}_{\text{Bias}(\hat{y})^2} + \underbrace{E[(\hat{y}-E[\hat{y}])^2]}_{\text{Var}(\hat{y})} + \underbrace{\sigma_\varepsilon^2}_{\text{Irreducible error}} \tag{4.1}$$

Accordingly, the out-of-sample MSE consists of a bias and variance term of the forecast and an irreducible error term, σ_ε^2. For the OLS estimator the bias term is zero. However, the OLS estimator does not necessarily minimize the out-of-sample error variance. By altering the estimated parameter values via shrinkage one introduces a bias into the forecast. However, if the reduction in the variance term exceeds the increase in (squared) bias, the out-of-sample error variance can be smaller. This is commonly referred to as the bias-variance trade-off. It is the foundation on why shrinkage can improve forecast accuracy (Hastie et al., 2009, pp. 223-228; G. James et al., 2023, pp. 31-34).

Linear Hybrid Shrinkage Let us now consider shrinkage within a specific method for forecast combination to further showcase how shrinkage can improve forecast accuracy. To this end, we introduce the linear shrinkage (LS) approach of Blanc and Setzer (2020)



(a) LS, i.e., $\boldsymbol{v} = (1, 1, 1, 1, 1, 1)'$ **(b)** LHS, i.e., $\boldsymbol{v} = (1, 0, 1, 1, 0, 0)'$

Figure 4.1. Illustration of forecast weight paths for the LS and LHS approach. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts are chosen randomly out of 24 once and are used throughout this thesis.

and its extension the linear hybrid shrinkage (LHS) approach of Schulz, Setzer, and Balla (2022). We chose these two methods for the illustration of shrinkage, because these are very accessible. They shrink the optimal weights towards equal weights linearly as a shrinkage parameter or intensity, λ, changes. However, within the linear hybrid shrinkage from Schulz et al. (2022) some weights are instead shrunken towards zero while the remaining subset of weights is shrunken towards the equal weights of the subset. The linear hybrid shrinkage weights are given by

$$\boldsymbol{\omega}^{LHS} = \lambda \frac{\boldsymbol{v}}{\mathbf{1}'\boldsymbol{v}} + (1 - \lambda)\boldsymbol{\omega}^{OW}, \tag{4.2}$$

with $\mathbf{1}$ being a $N \times 1$ vector of ones and the $N \times 1$ selection vector $\boldsymbol{v}$. The latter is used to select which forecasts are shrunk towards their corresponding equal weights and which are shrunk towards zero. For all vector elements that are one, the corresponding weights are shrunken linearly towards the respective equal weights, i.e., $1/\mathbf{1}'\boldsymbol{v}$. For all other vector elements that are zero, the weights are shrunken linearly towards zero. If $\boldsymbol{v}$ only consists of ones the LHS simplifies to LS used in Blanc and Setzer (2020). In order to have the transition between optimal and equal/zero weights, for the shrinkage parameter or intensity it holds that $\lambda \in [0, 1]$. If it is zero, the solution is identical to optimal weights. If it is one, some weights are identical to their corresponding equal weights while others are equal to zero. In general, the selection vector $\boldsymbol{v}$ can be determined by choice or, for example, based on the in-sample forecast accuracy (Blanc & Setzer, 2020; Schulz et al., 2022).

Figures 4.1(a) and 4.1(b) depict the weights (ordinate) of LS and LHS respectively for different values of λ (abscissa). Each colored line corresponds to the weight of a forecast for different values of λ. To put it differently, the colored lines are the path

each weight takes for all values of λ.[16] The underlying data is taken from the simulation study introduced in Section 3.2. Henceforth, we use the same data and the same six forecasts for all illustrations of the shrinkage methods within Chapter 4 unless indicated otherwise. The colors for all forecasts are the same for all illustrations.

Figure 4.1(a) illustrates how the linear shrinkage method operates. For $\lambda = 0$ weights are identical to the optimal weight of the forecast combination problem by J. M. Bates and Granger (1969). As the shrinkage parameter λ increases, all weights are shrunken simultaneously and linearly towards equal weights. For the LHS depicted in Figure 4.1(b) we selected forecasts based on their in-sample MSE values. The best three forecasts are shrunken towards equal weights (red, teal, and green) and the remaining three forecasts towards zero (pink, blue, and yellow).

Actual and Empirical Error Variance LS and LHS are straightforward shrinkage methods that showcase the general idea of shrinkage. Based on this, we can now further illustrate how shrinkage is beneficial. Consider the actual error variance of a combined forecast $\hat{y}_c$ for a given weight vector $\boldsymbol{w}$, i.e.,

$$\sigma_c^2(\boldsymbol{\omega}) = \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega}. \tag{4.3}$$

As discussed earlier, the true error covariance matrix $\boldsymbol{\Sigma}$ is, of course, unknown and has to be estimated. The corresponding empirical error variance is given by

$$\hat{\sigma}_c^2(\boldsymbol{\omega}) = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega}. \tag{4.4}$$

If we use simulated data for which we know the true error covariance matrix, we can analyze the actual and empirical error variance for different weight vectors $\boldsymbol{\omega}$. The most interesting weight vectors are of course the estimated weight vectors, $\widehat{\boldsymbol{\omega}}^{LHS}(\lambda)$, determined by Equation (4.2) with $\boldsymbol{v} = (1, 1, 1, 1, 1, 1)'$ (LS approach) for different values of λ. Figure 4.2 depicts the actual and empirical error variance, i.e., $\sigma_c^2(\widehat{\boldsymbol{\omega}}^{LHS}(\lambda))$ and $\hat{\sigma}_c^2(\widehat{\boldsymbol{\omega}}^{LHS}(\lambda))$ respectively, for the LS method.[17] The shrinkage parameter λ is on the abscissa. The ordinate shows the error variance, both the actual and empirical. The former is depicted by the blue colored line and the latter by the orange colored line. The smallest empirical error variance is, of course, given for $\lambda = 0$. Because then the weights from LS are identical to the (empirically) optimal weights of Equation (2.22), i.e., the weights that minimize the empirical error variance. The equal weights solution

[16] The figures we use that depict the weight paths are for example used in Hastie et al., 2009, pp. 57-73; Tibshirani, 1996

[17] Note that Blanc and Setzer (2020) derived an explicit decomposition of the expected squared error for the LS method. However, we introduce the concept of empirical and actual error variance, because it showcases the bias-variance trade-off, and it can be applied to any method that uses a hyperparameter (in the context of a simulation study), without the need of an explicit squared error decomposition for each individual method.

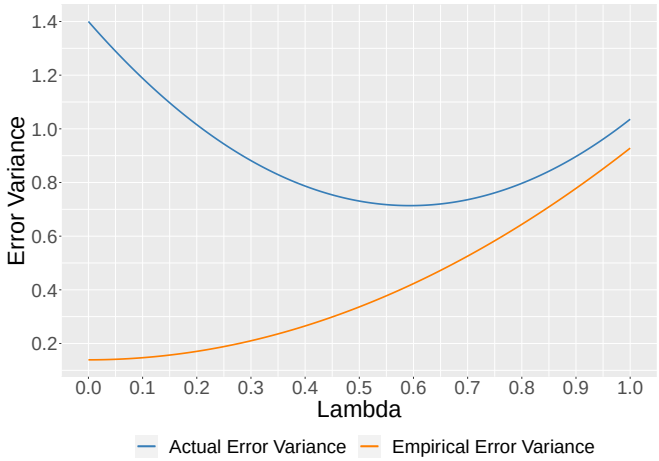


Figure 4.2. Illustration of the actual and empirical error variance for the LS method based on Fan et al. (2012). We use a different data set with $N = 24$ compared to Figure 4.1 for illustration purposes. The data was created on the basis of the simulation study of Section 3.2 for $CM1$, an error variance similarity $z = 0.5$ and no special group.

is given for $\lambda = 1$. The more weights are shrunken towards equal weights, the higher the empirical error variance gets. While the minimum of the empirical error variance is given for $\lambda = 0$, the actual error variance is minimized for about 0.56. While the empirical error variance of the combined forecast is non-decreasing for increasing values of λ, this is not true for the actual error variance. To put it differently, the weight vector that minimizes the actual error variance, which is based on the true but unknown covariance matrix of forecast errors, is in-between equal and optimal weights. In addition to that, Figure 4.2 also illustrates the forecast combination puzzle as although OW has the smallest empirical error variance, the actual error variance of the EW method is smaller in comparison.

The actual error variance in Figure 4.2 showcases the bias-variance trade-off of Equation (4.1). The solution of the forecast combination problem of Equation (2.22), i.e., OW, fits the in-sample data most accurately. In contrast, equal weights does not involve any estimation based on the in-sample data, i.e., the observed data is completely ignored. The OW method is unbiased, the EW is biased.[18]. For OW the higher variance term of the out-of-sample MSE decomposition in Equation (4.1) is a result of (over)fitting the in-sample data. Accordingly, it is smaller for EW than OW (Blanc & Setzer, 2020). For $\lambda = 0$ the LS method starts with an unbiased solution that has as larger variance term. It ends at $\lambda = 1$ with a biased solution that has a smaller variance term. On the way from $\lambda = 0$ to $\lambda = 1$, the bias term increases and the variance decreases. From the perspective of OW, if the increase in (squared) bias is more than compensated for by the reduction in variance, the resulting out-of-sample

[18]Note that OW is unbiased if the forecasts themselves are unbiased, recall Section 2.2.1

error variance or MSE is smaller than that of the OW method. Similarly, for EW, if the reduction in (squared) bias compensates the increase in the variance term, the error variance decreases. Between OW and EW there is (or can be) a λ value at which the bias increases is no longer compensated for by the reduction in the variance term or vice-versa, i.e., the minimum of the actual error variance. This value of λ minimizes the out-of-sample error variance of the LS method, i.e., has a superior out-of-sample forecast accuracy.

Shrinkage or regularization is build around this idea of the bias-variance trade-off. The weights are altered or shrunken into a certain direction or reference to reduce the out-of-sample error variance in terms of Equation (4.1) or similarly to reduce the actual error variance in terms of Equation (4.3) and Figure 4.2. As a result, there is an ongoing search for the best methods and shrinkage directions. In this thesis we will discuss and develop various methods to shrink forecast combination weights in order to trade bias and variance to reduce the out-of-sample error variance, MSE or forecast accuracy of the combined forecast. We start by taking a look at how Lasso-based shrinkage methods have already been used for forecast combination before we represent all those methods within a generalized framework and propose new shrinkage directions.

4.1.2 Lasso-based Shrinkage in the Forecast Combination Literature

The Least Absolute Shrinkage and Selection Operator (*Lasso*), was introduced by Tibshirani (1996). It imposes a L_1 constraint on the parameter vector in a regression problem. Because the Lasso-based method and other related shrinkage methods are based on vector norm, we briefly introduce these norms first. The L_q norm of a vector $\boldsymbol{\omega}$ for $q \geq 1$ is given by

$$\|\boldsymbol{\omega}\|_q = \left(\sum_{i=1}^{N} |\omega_i|^q \right)^{\frac{1}{q}}. \tag{4.5}$$

Accordingly, the L_1-norm ($q = 1$) is the sum of absolute weights, i.e.,

$$\|\boldsymbol{\omega}\|_1 = \sum_{i=1}^{N} |\omega_i|. \tag{4.6}$$

The L_2 norm ($q = 2$) is equivalent to the square root of the sum of squared vector elements, i.e.,

$$\|\boldsymbol{\omega}\|_2 = \left(\sum_{i=1}^{N} \omega_i^2 \right)^{\frac{1}{2}}. \tag{4.7}$$

Accordingly, $\|\boldsymbol{\omega}\|_2^2$ is the sum of squared vector elements (see e.g., Gentle, 2017, pp. 25-28). There is another (pseudo) norm L_0 that measures the sparsity of a vector, i.e., the number of non-zero elements. It is given as

$$\|\boldsymbol{\omega}\|_0 = \sum_{i=1}^{N} I(\omega_i \neq 0). \tag{4.8}$$

The function $I()$ is an indicator function that is equal to one if the condition is true, i.e., if the vector element is non-zero, and zero otherwise (see e.g., Lanza, Morigi, Selesnick, & Sgallari, 2023, pp. 13-14) Because we mostly use the L_1-norm, henceforth, we denote it by $\|\boldsymbol{\omega}\|$, i.e., we omit the subscript. For the L_0 and L_2 we will explicitly use $\|\boldsymbol{\omega}\|_0$ and $\|\boldsymbol{\omega}\|_2$.

Lasso The standard Lasso introduced by Tibshirani (1996) was already briefly mentioned in Section 2.4. It is defined within a regression framework by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \sum_{t=1}^{T} \left\| y_t - \boldsymbol{\omega}' \hat{\boldsymbol{y}}_t \right\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\omega}\| \leq \gamma \end{aligned} \tag{4.9}$$

with $\gamma \geq 0$. The objective function is the sum of squared residuals, i.e., it is equivalent to the MSE (without the constant term, see again Equation (2.2)). The constraint restricts how far the weights (or regression coefficients) as a whole can deviate from zero. It does that by constraining the sum of absolute weights to be smaller or equal to a shrinkage parameter γ. To put it differently, the weight vector $\boldsymbol{\omega}$ has a certain budget of deviation from zero. This budget can be distributed among all weights or only spend one weight depending on the objective function. If $\gamma = 0$ no deviation from zero is allowed, i.e., the only feasible solution is that all weights are zero. As γ increases, weights can become non-zero. For s sufficiently large value of γ which we denote as γ^* the constraint will no longer restrict the solution space. Accordingly, for $\gamma \geq \gamma^*$ the determined weights are identical to the OLS regression weights, i.e., the constraint itself becomes unnecessary in the sense that it does not affect the solution any more (Hastie et al., 2009, pp. 57-73; Tibshirani, 1996).

The Lasso can also be written as an unconstrained optimization problem. To this end, the inequality constraint is incorporated as a penalty into the objective function. This is called *Lagrangian(-multiplier) form* (Hastie et al., 2009, pp. 57-73; Tibshirani, 1996). The corresponding problem is given by

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \quad \sum_{t=1}^{T} \left\| y_t - \boldsymbol{\omega}' \hat{\boldsymbol{y}}_t \right\|_2^2 + \lambda \|\boldsymbol{\omega}\| \tag{4.10}$$

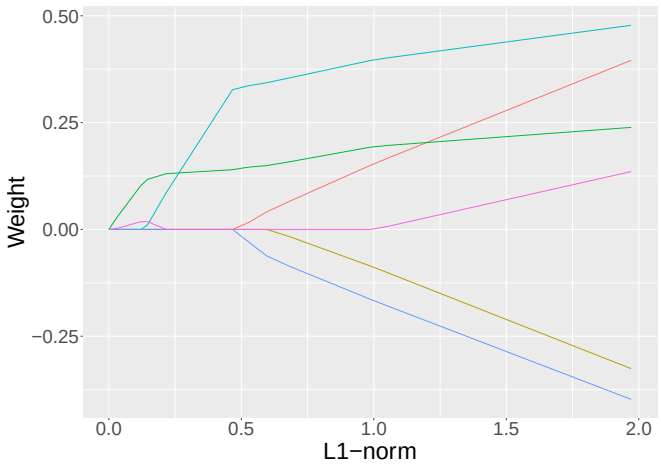


Figure 4.3. Illustration of forecast weight paths for the standard Lasso. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

The penalty term λ penalizes deviations from zero.[19] Similar to γ of Equation (4.9), for the penalty term λ it has to hold that $\lambda \geq 0$. However, while for $\gamma = 0$ all weights are zero, it is different for the Lagrangian form. Because λ is part of a penalty term instead of a constraint if $\lambda = 0$, i.e., no penalty, the determined weights are identical to the OLS weights. In contrast, for $\gamma = 0$ no deviation from zero is allowed. As the penalty term λ increases, i.e., as $\lambda \to \infty$, weights are shrunken towards zero. For a sufficiently large penalty term λ, i.e., λ^* all weights are zero (Hastie et al., 2009, pp. 57-73; Tibshirani, 1996).

The outstanding feature of the lasso both in form of Equations (4.9) and (4.10) is, that it shrinks weights towards zero and selects individual weights to exactly zero.[20] Figure 4.3 depicts the weights of six forecasts calculated for various values of λ/γ on the ordinate as colored paths. The abscissa shows the sum of the absolute values of the weights, i.e., $\|\boldsymbol{\omega}\|$ for the currently used γ or λ. By that, regardless of whether the constrained or unconstrained, Lagrangian form, is used, the figure has the same structure. We will henceforth use γ, but everything holds accordingly for λ. If $\gamma = 0$, i.e., $\|\boldsymbol{\omega}\| = 0$ all weights in Figure 4.3 are zero. As γ increases, i.e., more deviation from zero is allowed, two weights increase. The weights are depicted by the green and pink line. While the green weights increases further, the pink goes back to zero and the teal weight increases. For larger values of γ more and more weights deviated from zero and strive towards the unconstrained, OLS solution (blue, red, yellow, and pink). One can

[19] Note that we use λ if the Lagrangian form is used and γ if a constrained optimization problem is considered to distinguish between the two.

[20] The selection to exactly zero is because the constraint region defined by the L_1 constraint is, in two-dimensional terms, a diamond which corners occur where a variable is zero. See Hastie et al. (2009, pp. 68-73) and Tibshirani (1996) for more details.

also interpret Figure 4.3 from right to left. Starting from the OLS solution, the weights are shrunken and individually selected to zero as γ decreases.

If instead the L_2 norm is used within Equations (4.9) and (4.10), the problem is called *Ridge Regression*. It shrinks weights towards zero (and during that towards each other) but does not select individual weights to exactly zero (Diebold & Shin, 2019; Hastie et al., 2009, pp. 61-73; Tibshirani, 1996). In this thesis, however, we focus on the Lasso because it is well-established and widely used in the literature (Diebold & Shin, 2019; Radchenko et al., 2023; Roccazzella et al., 2022) on the one hand and the advantageous property of variable selection it provides on the other (recall Section 2.4 and see Wang, Kang, and Li (2022)). The latter can be particularly useful when only a few observations T but many forecasts N are available. In what follows we discuss how the Lasso has been used for forecast combination (Diebold & Shin, 2019; Hastie et al., 2009, pp. 57-73; G. James et al., 2023, pp. 229-253; Tibshirani, 1996).

eLasso and peLasso Diebold and Shin (2019) discuss multiple Lasso-based methods for forecast combination. Because they define their optimization problem both with the L_1-norm (Lasso) and L_2-norm (Ridge) we will at first present it in form of the L_q norm. However, later we will focus on the Lasso methods. For forecast combination Diebold and Shin (2019) minimize the squared residuals with a penalty term, i.e.,

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \quad \sum_{t=1}^{T} \left\| y_t - \boldsymbol{\omega}' \hat{\boldsymbol{y}}_t \right\|_2^2 + \lambda \|\boldsymbol{\omega}\|_q \tag{4.11}$$

If $q = 0$ the number of non-zero elements is penalized based on λ, i.e., the best subset of forecasts is selected and weights are estimated without any regularization. For $q = 2$ this optimization problem corresponds to the Ridge regression which, again, shrinks coefficients to zero but does not select a subset of forecasts to exactly zero. Both shrinkage and selection towards zero can be achieved by using $q = 1$ which is the Lasso regression (Diebold & Shin, 2019; Hastie et al., 2009, pp. 57-73; G. James et al., 2023, pp. 229-253; Tibshirani, 1996).

Based upon the fact that the equal weights forecasts oftentimes outperforms the theoretically optimal weights for forecast combination (see Section 2.3), the authors introduce shrinkage towards equal weights instead of zero. This can be achieved by using the following optimization problem

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \quad \sum_{t=1}^{T} \left\| y_t - \boldsymbol{\omega}' \hat{\boldsymbol{y}}_t \right\|_2^2 + \lambda \left\| \boldsymbol{\omega} - {}^1\!/_N \right\|_q \tag{4.12}$$

For $q = 2$ the N forecasts are shrunk towards the equal weights forecast. If the Lasso is used with $q = 1$, some forecasts are selected to have equal weights of ${}^1\!/_N$ while the weights of the remaining forecasts are shrunken towards it. They call those methods the

"egalitarian Ridge" (eRidge) and *"egalitarian Lasso" (eLasso)* (Diebold & Shin, 2019) respectively. To get a better intuition for the eLasso consider again Figure 4.3 but assume that the ordinate depicts $\omega_i - 1/N$ instead of ω_i. Then, a weight path for the eLasso can look similar to the weight path of the Lasso depicted in Figure 4.3. Accordingly, all forecasts have a weight of $1/N$ on the left-hand side of the figure and (oftentimes separately) start deviating from it for larger values of the L_1 norm on the abscissa.

Neither the eRidge nor the eLasso selects a subset of forecasts to zero (variable selection). The eLasso only selects forecasts to equal weights. Therefore, Diebold and Shin (2019) introduce the *"partially-egalitarian Lasso" (peLasso)* which corresponds to the following optimization problem:

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \quad \sum_{t=1}^{T} \left\| y_t - \boldsymbol{\omega}' \hat{\boldsymbol{y}}_t \right\|_2^2 + \lambda_1 \|\boldsymbol{\omega}\| + \lambda_2 \left\| \boldsymbol{\omega} - 1/\|\boldsymbol{\omega}\|_0 \right\| \tag{4.13}$$

The peLasso is designed to select towards zero and shrink towards equal weights, but not the original equal weights ($1/N$). Instead, it shrinks the weights towards the equal weights of the currently non-zero forecasts, i.e., $1/\|\boldsymbol{\omega}\|_0$. A problem with the peLasso is, however, that the optimization problem is difficult to solve, *"due to the discontinuity of the objective function at"* (Diebold & Shin, 2019) $\omega_i = 0$. Therefore, the authors propose a two-step procedure. First, a subset of forecasts is selected. This can be done by the standard Lasso method of Equation (4.11) with $q = 1$. Second, the forecasts with non-zero weights of step one are used within the eLasso of Equation (4.12) with $q = 1$.[21] As a consequence, the chosen forecasts are both shrunken and selected towards equal weights.

Henceforth, we will not consider the peLasso or an adapted version of it that includes the unity constraint within this thesis. In its current form as a two-step procedure, it is beyond the scope of this thesis. Our newly proposed forecast combination methods will be computed within one-step. Therefore, to fairly evaluate our methods to existing ones, we restrict ourselves exclusively on one-step procedures. Otherwise, if we apply the variable selection via a form of Lasso first and then apply the eLasso, we would also have to additionally evaluate all other methods after step one. To put it differently, we can also think about the pool of available forecast as a result of a variable selection strategy, i.e., a step one. However, that does not mean that we do not consider methods that perform a variable selection *simultaneously* in one step beside estimating weights, like for example the Lasso.

[21] Diebold and Shin (2019) also propose to use eRidge at this step.

Lasso towards Reference Weights and with Unity Constraint In Roccazzella et al. (2022) various approaches for constrained optimization with penalty terms are used. They define the optimization problem as a quadratic optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} + \lambda\|d(\boldsymbol{\omega}, \dot{\boldsymbol{\omega}})\| \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1 \end{aligned} \tag{4.14}$$

The first difference to the standard Lasso (as well as eLasso and peLasso) is, that they use the covariance matrix of forecast errors instead of the approximate objective function, i.e., the sum of squared residuals or MSE. By that, they follow the OW forecast combination problem by J. M. Bates and Granger (1969). Second, Roccazzella et al. (2022) also incorporate the unity constraint. The term $d(\boldsymbol{\omega}, \dot{\boldsymbol{\omega}})$ is a measure of diversion between the weight vector $\boldsymbol{\omega}$ and a reference or *prior* weight vector $\dot{\boldsymbol{\omega}}$. As a diversion measure Roccazzella et al. (2022) use, inter alia, an elastic net. It is a hybrid between the Lasso, L_1-norm, and Ridge, L_2-norm, and is given by

$$a\|\boldsymbol{\omega} - \dot{\boldsymbol{\omega}}\| + (1-a)\|\boldsymbol{\omega} - \dot{\boldsymbol{\omega}}\|^2 \tag{4.15}$$

with $a \in [0, 1]$. If $a = 0$ the optimization problem is solved under a Ridge penalty. If $a = 1$ instead, the problem has a Lasso penalty. As a reference weight $\dot{\boldsymbol{\omega}}$ they use both the equal weights forecasts and the inverse-loss (IL) weighted average. The latter is given by

$$\omega_i^{IL} = \left(\sigma_i^2 \sum_{j=1}^{N} \frac{1}{\sigma_j^2}\right)^{-1} \qquad \forall\ i = 1, \ldots, N. \tag{4.16}$$

For the inverse-loss weighted average, the smaller the error variance or forecast accuracy of a forecast is, the larger the corresponding weight.

Double Shrinkage via Weighted Least Squares Lasso Recently, Liu, Hao, and Wang (2023) proposed a weighted least squares Lasso approach (WLS-Lasso) that also shrinks towards equal weights and zero. Shrinkage towards equal weights is achieved by including a weighting factor into the regression-based objective function, i.e., the sum of squared residuals. This weighting factor is determined by the sum of squared residuals of the equal weights forecasts as well as a penalty term λ_1. For $\lambda_1 = 0$ the approach corresponds to the regression-based implementation of the forecast combination problem by J. M. Bates and Granger (1969). The larger λ_1 gets the closer the solution is forced towards equal weights.

In addition to that Liu et al. (2023) included a penalty term into the objective function that shrinks towards zero. To this end one can use a Lasso ($q = 1$) as well as Ridge ($q = 2$) penalty, i.e., $\lambda_2\|\boldsymbol{\omega}\|_q$ as in Equation (4.11), or an elastic net penalty, see Equa-

tion (4.15), into the objective function to additionally shrink the forecast combination weights towards zero. Although their approach is interesting and would fit into this thesis, due to its very recent publication it can not be taken into consideration in this thesis. Nevertheless, an evaluation of it in comparison to the proposed methods and the inclusion of it into the later proposed unified framework has to be considered in the future.

Lasso-based Methods in Forecast Combination Lastly, in independent and concurrent work we and Radchenko et al. (2023) considered the forecast combination problem with a L_1 constraint as a quadratic optimization problem without using a Lagrangian relaxation. At its core, this approach is the adaptation of the standard Lasso as a quadratic optimization problem within the context of forecast combination. First, the approach is based on the error variance-covariance matrix like the OW forecast combination problem by J. M. Bates and Granger (1969). In contrast, the standard Lasso, Equation (4.10), as well as the forecast combination adaptations in form of the eLasso or peLasso, Equations (4.12) and (4.13), use the approximated objective function in form of the squared residuals instead. Second, the approach incorporates the unity constraint, again in contrast to the standard Lasso, eLasso and peLasso, which ensures the unbiasedness of the combined forecast. Moreover, as we discussed previously constraining the solution space by the inclusion of the unity constraint can improve forecast accuracy. Third, the approach imposes a L_1 constraint directly and not a relaxation within the objective function in contrast to Roccazzella et al. (2022), see Equation (4.14).

Henceforth, we will reformulate a selection of the approaches by Diebold and Shin (2019) and Roccazzella et al. (2022) into a general framework. To this end and in accordance to Radchenko et al. (2023) we use a quadratic optimization problem instead of the regression framework because we base our analysis on the original forecast combination problem by J. M. Bates and Granger (1969). Additionally, the squared residuals are only an approximation of the error variance of the combined forecast. We follow the original idea of J. M. Bates and Granger (1969) and assume unbiased forecasts. Therefore, we will reformulate the approaches to include the unity constraint. Additionally, we formulate and implement the optimization problems with actual constraints instead of a Lagrangian relaxation.

In what follows, we will incorporate the L_1 constraint and extension, like for example the eLasso, into a unified forecast combination framework. Moreover, we will propose new shrinkage directions inspired by the eLasso and LHS.

4.2 A Unified Framework for Lasso-based Forecast Combination Methods

In this section we provide a general, unified framework for the incorporation of the L_1 constraints into forecast combination. To this end, we will transform the previously discussed approaches of Section 4.1 into a quadratic optimization problem as it was proposed by J. M. Bates and Granger (1969). First, we present them in terms of the L_1-norm constraint. Second, we transform the constraints into linear constraints. If the L_1 constraints are included into the optimization problem directly, it is a non-linear optimization problem with non-differentiable constraints (see e.g., Luenberger & Ye, 2016, Chapter 1, 2, 11 particularly pp. 321-323; Schmidt, 2005).

The forecast combination methods as a quadratic optimization problem with linear inequality constraints is given by

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \frac{1}{2}\boldsymbol{x}'\boldsymbol{D}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}'\boldsymbol{x} \geq \boldsymbol{b} \end{aligned} \tag{4.17}$$

The elements in the vector $\boldsymbol{x}$ are N variables, i.e., weights in the context of forecast combination. $\boldsymbol{D}$ is a $N \times N$ *symmetric* matrix. Additionally, $\boldsymbol{A}$ is a $N \times M$ matrix that defines M different constraints. The corresponding right-hand side of the constraint is defined by the $N \times 1$ vector $\boldsymbol{b}$. We solve the optimization problems in the statistical software R (R Core Team, 2022) with the R package *quadprog* (Turlach, Weingessel, & Moler, 2019). As a result we use the method of Goldfarb and Idnani (1982, 1983) to solve the problem. For this method it is required that $\boldsymbol{D}$ is positive definite. However, upon the implementation of the Lasso-based methods introduced in Section 4.1, the matrix is not guaranteed to be positive definite. Therefore, we follow Radchenko et al. (2023) and use the function *nearPD* from the R package *Matrix* (D. Bates et al., 2022). The function computes the nearest positive definite matrix based on the algorithm in Higham (2002) (for more details see D. Bates et al., 2022).

The fact that Equation (4.17) only has linear inequality constraint (greater or equal to) is not a problem as it is easy to incorporate both inequality constraints (smaller or equal to) and equality constraints. For example, consider the OW forecast combination problem of J. M. Bates and Granger (1969) of Equation (2.22). To reformulate the unity constraint, i.e., $\sum_{i=1}^{N} \omega_i = 1$, into the form of the optimization problem of Equation (4.17), we introduce the following two constraints:

$$\sum_{i=1}^{N} \omega_i \geq 1 \tag{4.18}$$

$$\sum_{i=1}^{N} \omega_i \leq 1 \tag{4.19}$$

We then multiply the second constraint with negative one, i.e.,

$$\sum_{i=1}^{N} -\omega_i \geq -1. \tag{4.20}$$

By that, we reformulated the forecast combination problem with the unity constraint such that it has the same form as Equation (4.17) (see e.g., Hurlbert, 2010, pp. 7-9).

The remainder of this section is organized as follows. In Section 4.2.1 we introduce how shrinkage towards zero with an L_1 constraint can be implemented with linear constraints and analyze how it effects the weights. Thereafter, in Section 4.2.2 we generalize this approach. First for shrinkage towards any fixed value κ and then towards any prior weights vector $\boldsymbol{\omega}$ which nests all previously considered shrinkage directions. By that, we gradually construct the unified framework that incorporates all considered L_1 constraints. Lastly, in Section 4.2.3 we propose to use a new shrinkage direction for the L_1 constraint that is inspired by the LHS, eLasso and peLasso.

4.2.1 The L_1 Constraint for Shrinkage and Selection towards Zero

Another way to look at forecast combination is from the perspective of portfolio selection or optimization introduced by Markowitz (1952). The underlying optimization problem from portfolio selection is identical to the forecast combination problem of Equation (2.22). Instead of combining forecasts, in portfolio selection assets are combined into a portfolio. Instead of the covariance matrix of forecasts errors, portfolio selection uses the covariance matrix of assets within the objective function. In the portfolio context, weights correspond to a proportion of the budget an investor invests in one asset. The weights assigned to the assets also have to sum up to unity, which corresponds to a budget constraint that, however, implies that all the budget has to be spent. Weights greater one or smaller zero correspond to long and short positions respectively. They are financial instruments or tools that allow investors to bet on increasing or decreasing asset prices in the future (Arratia, 2014, pp. 19-20, 239-243; Fan et al., 2012; Markowitz, 1952). If both short and long positions are forbidden, a no-short-sale constraint is imposed, i.e., $\omega_i \geq 0$ for all assets $i = 1, \ldots, N$. Then the problem is identical to the PW approach of Equation (2.30). Fan et al. (2012) implemented the L_1 constraint to bridge the gap between the no-short-sale portfolio and the unconstrained portfolio optimization problem which allows for arbitrary large short (and long) positions.

Inspired by Fan et al. (2012), Radchenko et al. (2023) published an article that adopted the approach to forecast combination. For that, they used the framework of a

constrained quadratic optimization problem, as proposed by J. M. Bates and Granger (1969), that we want to use for all L_1 constrained methods. The problem is defined by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \|\boldsymbol{w}\| \leq \gamma \end{aligned} \tag{4.21}$$

Because of the L_1 constraint this problem shrinks and selects towards zero. In comparison to the standard Lasso of Equation (4.9), the weights can not all be zero. This would lead to a violation of the unity constraint. At least one weight has to be non-zero to fulfill the unity constraint. Accordingly, in contrast to the standard Lasso, the smallest feasible γ value for the optimization problem of Equation (4.21) can not be zero. Moreover, all values $\gamma < 1$ result in an infeasible optimization problem. It has to hold that

$$\gamma \in [1, \infty]. \tag{4.22}$$

If $\gamma = 1$ the sum of absolute weights has to be smaller or equal to one. In conjunction with the unity constraint, this implies that only non-negative weights can provide a feasible solution. Thus, if $\gamma = 1$ the problem is identical to the PW approach. As γ increases a certain amount of negative weights and, consequently, weights greater one are allowed. As $\gamma \to \infty$ there is, again, a certain value γ^*. For any value of $\gamma > \gamma^*$ the L_1 norm does not constrain the solution space anymore, i.e., the solution is identical to the OW approach. It holds that $\gamma^* = \|\boldsymbol{\omega}^{OW}\|$, i.e., γ^* is the sum of absolute weights of the OW solution. By that, the L_1 constraint bridges the gap between the positive and optimal weights approach (PW and OW). It allows for solutions in-between these two optimization problems, i.e., forecast combination approaches. Instead of either only allowing positive (and zero) weights or not constraining negative weights at all, it allows for a certain amount of negativity within a solution. For a given solution $\boldsymbol{\omega}$, the amount of positive and negative weights is

$$w^{pos} = \frac{(\|\boldsymbol{w}\|_1 + 1)}{2} \qquad \text{and} \qquad w^{neg} = \frac{(\|\boldsymbol{w}\|_1 - 1)}{2}. \tag{4.23}$$

It holds that $w^{pos} + w^{neg} = \|\boldsymbol{w}\|_1$ as well as $w^{pos} - w^{neg} = 1$ (Fan et al., 2012).

Figure 4.4 depicts the weight paths for the same data used to depict the weight path of the standard Lasso in Figure 4.3. The ordinate shows the weights of the different forecasts, differentiated by colors. The forecasts have the same colors as in Figure 4.4. However, the abscissa in Figure 4.4 shows the γ value and not the L_1 norm of the weight vector. There are two reasons for that. First, by using the generalized constrained quadratic optimization framework, i.e., not the Lagrangian form with the penalty term

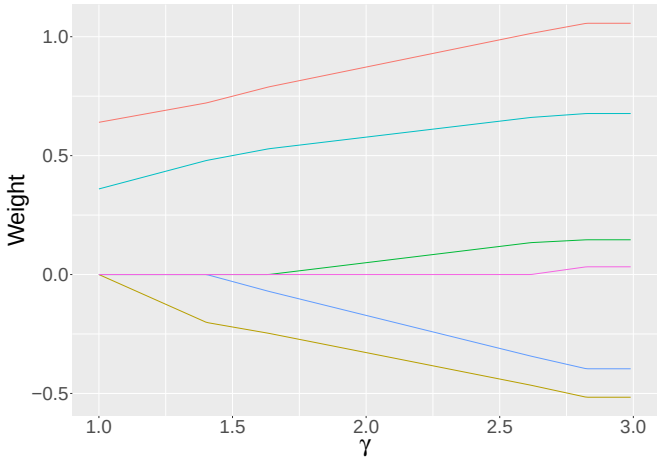


Figure 4.4. Illustration of forecast weight paths for the forecast combination problem with a L_1 constraint of Equation (4.21) for different γ-values. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

λ, there is no need to use the L_1 norm instead of γ as we only use γ and do not have to find a common abscissa that is identical regardless of which parameter (γ or λ) is used. Second, by using γ, we can illustrate that the solution does not change for $\gamma > \gamma^*$. This is not possible if we instead use the L_1 constraint on the abscissa, because it is identical for all solutions calculated with $\gamma \geq \gamma^*$. Figure 4.4 depicts the smallest feasible value of the shrinkage parameter $\gamma = 1$ on the left side. At that point, two weights are non-zero (red and teal). Again, this solution corresponds to the PW approach. As γ increases the two non-zero weights increase further. To offset this increase and fulfill the unity constraint, another weight (yellow) deviates from zero and becomes negative. The larger γ gets the more negativity is introduced into the solution overall. Furthermore, more weights deviate from zero (blue, green, and pink). For all γ values that are greater $\gamma^* \approx 2.8$ the solution is identical to the OW solution.

The weight paths give an example of how the method shrinks and selects weights. However, note that this is only exemplary and general conclusion of the weights paths can not be drawn from this. For example, based on Figure 4.4 one may conclude that after a weight deviates from zero, it is non-decreasing if it is positive and non-increasing if it is negative. However, this is not the case in general. To showcase this, we present in Figure A.1 Appendix A weight paths for, among others, the optimization problem Equation (4.21). Note that for those figures a different data set is used. It is based on the same simulation study scenario, i.e., $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. We include all 24 forecasts in Figure A.1.

Linear Constraints for the L_1-norm In order to solve the problem of Equation (4.21) we need to transform the L_1 norm into linear constraints in the form of the optimization problem of Equation (4.17). To this end, one can use two non-negative variables $\omega_i^+ = \omega_i$ if $\omega_i \geq 0$ and $\omega_i^- = \omega_i$ if $\omega_i < 0$ for each forecast or weight $\omega_i \; \forall \; i = 1, \ldots, N$ (see e.g., Schmidt, 2005). It holds that

$$\omega_i = \omega_i^+ - \omega_i^- \qquad \forall \; i = 1, \ldots, N. \tag{4.24}$$

If $\omega_i^+ > 0$ and, accordingly, $\omega_i^- = 0$, the weight ω_i is positive. If instead, $\omega_i^- > 0$ and $\omega_i^+ = 0$, the weight ω_i is negative. As a consequence, we have to define a new error covariance matrix

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \end{pmatrix}, \tag{4.25}$$

and weight vector

$$\tilde{\boldsymbol{w}} = (\omega_1^+, \omega_2^+, \ldots, \omega_N^+, \omega_1^-, \omega_2^-, \ldots, \omega_N^-)'. \tag{4.26}$$

Then we can represent the problem of Equation (4.21) by the optimization problem

$$\begin{aligned} \underset{\tilde{\boldsymbol{\omega}}}{\text{minimize}} \quad & \tilde{\boldsymbol{\omega}}' \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\omega}} \\ \text{subject to} \quad & \tilde{\boldsymbol{\omega}}' \tilde{\mathbf{1}} = 1, \\ & \tilde{\boldsymbol{\omega}}' \mathbf{1} = 1, \\ & \tilde{\omega}_i^+ \geq 0 \qquad \forall \; i = 1, \ldots, N, \\ & \tilde{\omega}_i^- \geq 0 \qquad \forall \; i = 1, \ldots, N \end{aligned} \tag{4.27}$$

with $\tilde{\mathbf{1}}$ being a $2N \times 1$ vector where the first N elements are 1 and the last N are -1. $\tilde{\boldsymbol{\omega}}' \tilde{\mathbf{1}}$ computes the sum over all weights, recall Equations (4.24) and (4.26), and restricts it to be equal to one (unity constraint). In contrast, $\mathbf{1}$ is a $2N \times 1$ vector of positive ones. By that, the constraint $\tilde{\boldsymbol{\omega}}' \mathbf{1}$ sums up both ω_i^+ and $\omega_i^- \; \forall \; i = 1, \ldots, N$ and thus $\tilde{\boldsymbol{\omega}}' \mathbf{1} \leq \gamma$ corresponds to the constraint that the sum of absolute weights is not greater γ (Schmidt, 2005). Accordingly, we can adapt and solve the Lasso with a unity constraint for forecast combination. However, we can further generalize the optimization problem to derive a unified framework that can incorporate different L_1 constraints or rather shrinkage directions.

4.2.2 A Generalized Approach for Shrinkage and Selection based on the L_1 Constraint

To adapt the eLasso from Diebold and Shin (2019) depicted in Equation (4.12) into the general optimization framework (including the unity constraint), we can derive more general linear constraints that introduce the L_1 norm into the optimization problem. For now, we assume a fixed value to which we shrink towards in Section 4.2.2.1. In Roccazzella et al. (2022) weights are also shrunken and selected towards a prior weights vector, $\dot{\boldsymbol{\omega}}$. Our unified framework of the forecast combination problem with the L_1 constraint also includes shrinkage towards prior weights which we will show in Section 4.2.2.2.

4.2.2.1 Shrinkage Towards a Fixed Value

The Lasso with a unity constraint presented in Section 4.2.1 and the eLasso briefly introduced in Section 4.1 both shrink towards a fixed value, zero and equal weights respectively. We can generalize by using any fixed values $\kappa \in \mathbb{R}$. The corresponding optimization problem is given by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \|\boldsymbol{w} - \kappa\| \leq \gamma_\kappa \end{aligned} \tag{4.28}$$

If $\kappa = 0$, this problem corresponds to the forecast combination problem with an L_1 constraint of Equation (4.21), i.e., weights are shrunken and selected towards zero. If instead $\kappa = 1/N$ weights are both shrunken and selected to equal weights, as intended by the eLasso. Henceforth, we will refer to these problems as $L_1(\kappa)$. Note that we also introduced a subscript κ for the shrinkage parameter γ, i.e., γ_κ, to clarify the affiliation to the respective constraint.

In general, $L_1(\kappa)$ can be used with any value of $\kappa \in \mathbb{R}$. Values of κ beside zero and $1/N$, however, are less intuitive, for example $\kappa = 2/N$ or if it has negative values. In the latter case, weights are shrunken and selected towards a negative weight. Depending on the value of κ, the feasible values for γ_κ are different. For example, if $\kappa = 0$, the smallest feasible value of γ is one. For the eLasso with a unity constraint, i.e., $\kappa = 1/N$, a value of zero for $\gamma_{1/N}$ fulfills the unity constraint as the resulting solution is equal weights. Instead, if for example $N \neq 10, \kappa = 0.1$ and $\gamma_{0.1} = 0$, the sum of absolute weights can not deviate from 0.1 and the unity constraint is violated. Accordingly, we have to derive an interval for the feasible values of γ_κ based on κ.

Accordingly, the smallest feasible value for γ_κ is given by

$$\gamma_\kappa \in \big[|1 - N\kappa|, \infty\big) \tag{4.29}$$

Note that we use the absolute value to cover both cases of $N\kappa$ greater and smaller one as a result of $\kappa \in \mathbb{R}$. The basic idea behind the smallest feasible value shown in Equation (4.29) is that we assume that all weights are equal κ, i.e., $\omega_i = \kappa \; \forall \; i = 1, \ldots, N$. In that case, we determine how much at least one weight has to deviate from this solution such that the unity constraint is fulfilled, i.e., weights sum up to unity.

Linearization for the Generalized L_1 Constraint Recall that the optimization problem introduced in Equation (4.27) of Section 4.2.1 incorporates the L_1 constraint directly with the introduction of two non-negative variables that correspond to the positive part and negative part of ω_i. However, to the best of our knowledge one can not implement shrinkage towards other values for κ than zero using this reformulation. Therefore, we use a different approach to reformulate the generalized optimization problem $L_1(\kappa)$ in terms of linear constraints.[22] To this end, we define

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{4.30}$$

where $\mathbf{0}_{N \times N}$ is a matrix that contains only zeros. By that, we introduce N additional variables $u_i \in \mathbb{R} \; \forall \; i = 1, \ldots, N$ that have no effect on the objective function.[23] The adjusted weight vector is given by

$$\tilde{\boldsymbol{\omega}} = (\omega_1, \omega_2, \ldots, \omega_N, u_1, u_2, \ldots, u_N)'. \tag{4.31}$$

Finally, we define the optimization problem as

$$\begin{aligned}
\underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \tilde{\boldsymbol{\omega}}' \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\omega}} && && \text{(4.32a)} \\
\text{subject to} \quad & \sum_{i=1}^{N} \omega_i = 1, && && \text{(4.32b)} \\
& \sum_{i=1}^{N} u_i \leq \gamma_\kappa, && && \text{(4.32c)} \\
& \omega_i - \kappa \leq u_i && \forall \; i = 1, \ldots, N, && \text{(4.32d)} \\
& \kappa - \omega_i \leq u_i && \forall \; i = 1, \ldots, N && \text{(4.32e)}
\end{aligned}$$

[22] For the sake of clarity, we do not imply that the following core optimization problem has not been used in other circumstances. It is a straightforward formulation of the optimization problem that we create to solve our problem.

[23] For the sake of completeness note that although we do not define u_i to be non-negative, it will always be in a given solution of the optimization problem. Otherwise, the optimization problem is infeasible, see Equations (4.33) and (4.34).

For now, assume that the additional variables u_i are given. For any weight $\omega_i \; \forall \; i = 1, \ldots, N$ we can rewrite Equations (4.32d) and (4.32e) as

$$\omega_i \leq u_i + \kappa, \tag{4.33}$$

$$\omega_i \geq -u_i + \kappa. \tag{4.34}$$

By that, it becomes clear that the constraints depicted in Equations (4.32d) and (4.32e) are bounds for each weight. To illustrate the bounds the following figure depicts the interval of feasible values around $\kappa = 0$ for $N = 2$ exemplary. Assume that it holds that $u_1 = 1.5$, $u_2 = 0.5$ and $\gamma_0 = 2$.

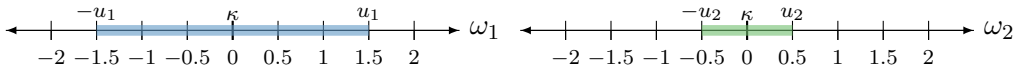


Figure 4.5. Illustration of the bounds imposed by u_i around each weight ω_i for $\kappa = 0$ with $u_1 = 1.5$, $u_2 = 0.5$ and $\gamma_0 = 2$.

The left number line in Figure 4.5 shows the feasible values for ω_1 while the right number line depicts them for ω_2. If $\kappa = 0$, it has to hold that $\omega_i \in [-u_i, u_i] \; \forall \; i = 1, \ldots, N$, see Figure 4.5. The feasible interval for ω_1 depicted as the blue rectangle or line is larger than the interval for ω_2 (green rectangle). As one can see u_i defines how much each weight can deviate from zero. Accordingly, if $\kappa \neq 0$, u_i defines how much each weight can deviate from κ. To put it differently, for $\kappa \neq 0$ the feasible interval for each weight ω_i is given by $[-u_i + \kappa, u_i + \kappa]$. Because the interval is symmetrical, it corresponds to constraining the absolute value of the difference between ω_i and κ, i.e., $|\omega_i - \kappa| \leq u_i$.

In summary, the constraints in of Equations (4.32d) and (4.32e) restrict the absolute difference between each weight ω_i and κ individually. However, the original L_1 constraint restricts the weight vector $\boldsymbol{\omega}$ as a whole. In order to achieve that we use the constraint in Equation (4.32c). It restricts the sum of all u_i to be smaller or equal to γ_κ. By that, it connects all individual constraints of Equations (4.32d) and (4.32e) that restrict the amount of deviation from κ by introducing a global budget for them.

Lastly, we have to again consider the variables u_i. They can be set to any value by the algorithm that solves the optimization problem, because they do not affect the objective function, see again Equations (4.30) and (4.31). If the algorithm sets ω_i to a specific value, it also has to change u_i to fulfill the corresponding constraints of Equations (4.32d) and (4.32e). If $\gamma < \gamma^*$, i.e., the γ still constraints the solution space, it holds that $u_i = |w_i - \kappa|$.

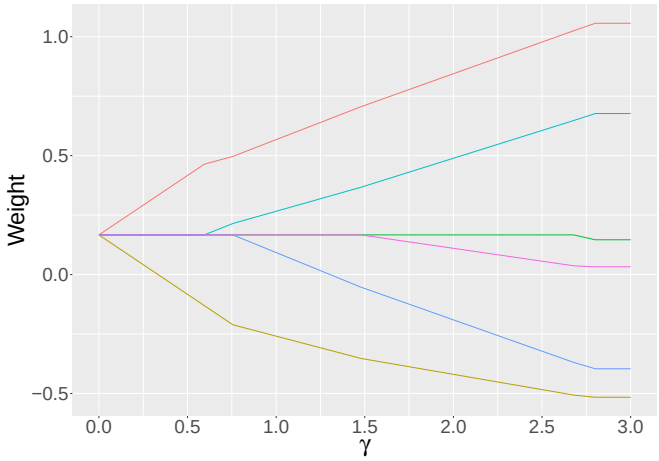
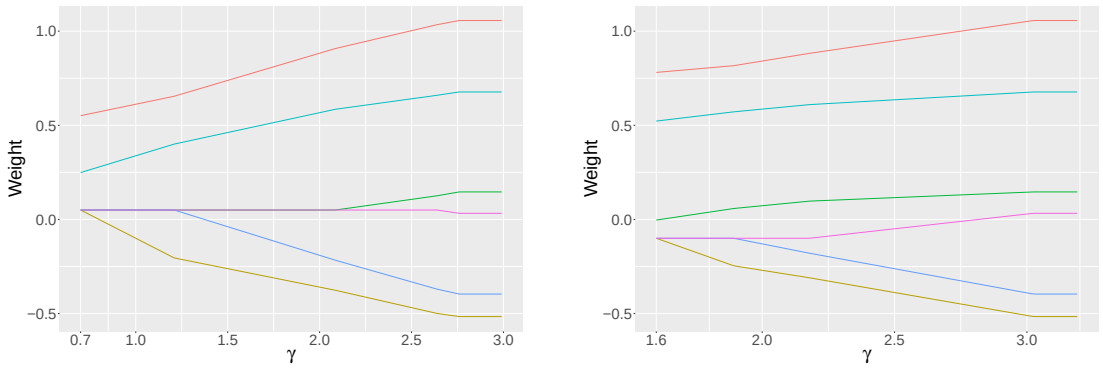


Figure 4.6. Illustration of forecast weight paths for the egalitarian Lasso with unity constraint for different γ-values. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

Weight Paths for the Generalized L_1 Constraint Let us now consider the solutions of the more general optimization problem of Equation (4.32). For $\kappa = 0$, the solution path is identical to the optimization problem defined in Equation (4.21), see again Figure 4.4. Figure 4.6 depicts the weights for $L_1(1/N)$, i.e., the equivalent to the eLasso with a unity constraint. The weight of each forecast is on the ordinate and the parameter $\gamma_{1/N}$ is depicted on the abscissa. For $\gamma_{1/N} = 0$, the only feasible solution is equal weights, i.e., $\omega_i = 1/N \; \forall \; i = 1, \ldots, N$. As $\gamma_{1/N}$ increases, weights deviate from equal weights. However, if one weight increases (red line), another has to decrease (yellow) to ensure that the unity constraint is fulfilled. As $\gamma_{1/N}$ is increased further, more weights deviate from equal weights (teal, blue, purple, green). Additionally, the deviation of weights that are already unequal to equal weights increases further (see e.g., red and yellow). To put it differently, if we look at Figure 4.6 from right to left, weights are shrunken and selected towards equal weights.

Figure 4.7 depicts the weight paths for both $\kappa = 0.05$ in Figure 4.7(a) and $\kappa = -0.1$ in Figure 4.7(b). The smallest feasible γ values are given on the left-hand side of both figures. If $\kappa = 0.05$ not all weights can be exactly equal to it, due to the unity constraint. Thus, as one can see in Figure 4.7(a), two weights (red and teal) deviate from κ by overall $|1 - 6 \cdot 0.05| = 0.7$, i.e., the smallest feasible $\gamma_{0.05}$ value determined by Equation (4.29). Thereafter, the weight path is in general similar to $\kappa = 0$ (Figure 4.4) and $\kappa = 1/N$ (Figure 4.6). The larger γ gets, weights that were previously equal to κ deviate (e.g., yellow and blue) and weights that are already different from κ change further (e.g., red and teal). To put it differently, from right to left, weights are shrunk and selected towards 0.05 under the obligation that the unity constraint is fulfilled. If



(a) $\kappa = 0.05$ **(b)** $\kappa = -0.1$

Figure 4.7. Illustration of forecast weight paths for the egalitarian Lasso with unity constraint and for different γ_κ and κ-values. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

$\kappa = -0.1$, as depicted in Figure 4.7(b), the weights path is, in general, similar. However, to ensure a feasible solution, the smallest feasible $\gamma_{-0.1}$ value has to be much larger with 1.6. Furthermore, at this point, three weights (red, teal, and green) are already different from κ. Thereafter, again, weights deviate from κ (yellow, blue, purple) and change further (red, teal, green) as $\gamma_{-0.1}$ increases. For both Figures 4.7(a) and 4.7(b), at a certain value γ^*, the weights are identical to the optimal weights solution.

Although, at first glance, choosing a different value for κ than zero or equal weights seems more arbitrary, it basically, shrinks weights towards a solution where a subset of weights is *identical*. The remaining weights are then *individual* to offset the identical weights such that the unity constraint is fulfilled. Considering again Figure 4.7(a) where $\kappa = 0.05$, for $\gamma_{0.05} = 0.7$ it has a resemblance to the positive weights solution depicted in Figure 4.6 where $\kappa = 0$ for $\gamma_0 = 1$. In both figures, all but two weights (red and teal) are identical. If $\kappa = 0$, they are identical with a weight of zero. In contrast, if $\kappa = 0.05$, they are identical at 0.05. To put it differently, by choosing $\kappa \neq 0$ one introduces a certain degree and type of contribution that each forecast has for the combined forecasts.[24] The weights are shrunk and selected towards that minimum contribution as γ decreases, i.e., from right to left in Figure 4.7.

A Note on the peLasso Before we move on to shrinkage towards prior weights in Section 4.2.2.2 we briefly want to provide a note on the peLasso. As discussed earlier in Section 4.1.2, as far as we are aware there is no one-step procedure for the peLasso. However, based on our unified framework for the L_1 constraint, we can provide a one-

[24] Note that degree refers to the overall magnitude of the weights and type to whether the contribution is positive or negative.

step procedure. The resulting problem is a mixed integer quadratic problem which is different from the quadratic optimization problem with linear constraint and real variables that we consider in this thesis. Therefore, it is beyond the scope of this thesis and we leave it for future research to implement our provided one-step procedure for the peLasso. To this end, we briefly introduce it in Appendix A. In Section 4.2.3 we will introduce a shrinkage direction that is inspired by the one-step peLasso (and LHS).

4.2.2.2 Shrinkage towards Prior Weights

In the previous Section 4.2.2.1 we used a fixed value κ to which weights are shrunk and selected. However, Roccazzella et al. (2022) also shrink and select weights towards some prior weights, e.g., the inverse-loss weighted average see again Equation (4.16). The generalization for the introduction of the L_1 constraint into the forecast combination problem presented in Section 4.2.2.1 can be easily adapted to this scenario and, thereby, we can provide a unified framework for the use of L_1 constraints for forecast combination. The optimization problem is given by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \|\boldsymbol{w} - \dot{\boldsymbol{\omega}}\| \leq \gamma_{\dot{\boldsymbol{\omega}}} \end{aligned} \tag{4.35}$$

We will refer to this problem as $L_1(\dot{\boldsymbol{\omega}})$. As we will show throughout the remaining part of Section 4.2 this is the unified framework or optimization problem that nests the considered L_1 methods and more. This is the second main contribution of this thesis.

In order to solve Equation (4.35), we can, similar to Equation (4.32), rewrite the optimization problem with linear constraints:

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \quad \tilde{\boldsymbol{\omega}}'\widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\omega}} \tag{4.36a}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \omega_i = 1, \tag{4.36b}$$

$$\sum_{i=1}^{N} u_i \leq \gamma_{\dot{\boldsymbol{\omega}}}, \tag{4.36c}$$

$$\omega_i - \dot{\omega}_i \leq u_i \qquad \forall\ i = 1, \ldots, N, \tag{4.36d}$$

$$\dot{\omega}_i - \omega_i \leq u_i \qquad \forall\ i = 1, \ldots, N \tag{4.36e}$$

The only difference, to Equation (4.32) is that the fixed value κ is replaced by the prior weight $\dot{\omega}_i$ within Equations (4.36d) and (4.36e). Similar to Figure 4.5 where κ was fixed, the feasible values of weights when prior weights are used are depicted exemplary for $N = 2$ in Figure 4.8. The prior weights are assumed to be $\dot{\omega}_1 = 0.25$

Figure 4.8. Illustration of the bounds imposed by u_i around each weight ω_i depending on the prior weight $\dot{\omega}_i$.

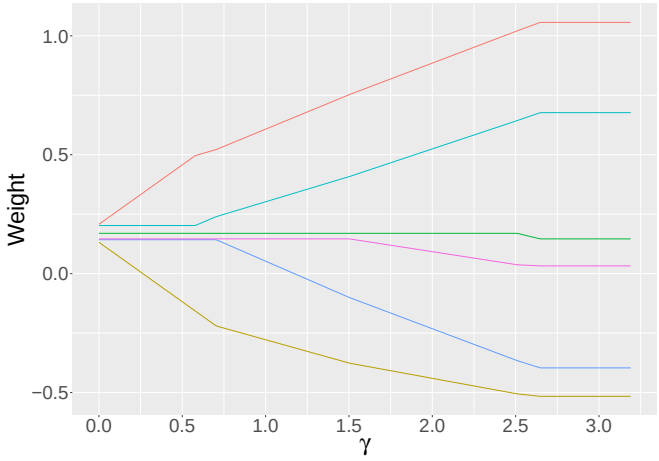


Figure 4.9. Illustration of forecast weight paths for the Lasso with unity constraint and prior weights for different $\gamma_{\dot{\boldsymbol{\omega}}}$. The prior weights are the inverse-loss weighed average, $\boldsymbol{\omega}^{IL}$. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

and $\dot{\omega}_2 = 0.75$. The blue and green rectangles depict the feasible intervals for each weight, i.e., $[-u_i + \dot{\omega}_i, u_i + \dot{\omega}_i]$. Within this example $\gamma_{\dot{\boldsymbol{\omega}}} = 1.5$. Both the first and second weight can deviate from their corresponding prior weights in both directions by about $u_1 = u_2 = 1$.

Figure 4.9 depicts weight paths for $L_1(\dot{\boldsymbol{\omega}})$ of Equation (4.35). For the prior weights we use $\boldsymbol{\omega}^{IL}$ of Equation (4.16). Recall, that for a fixed value of κ, a set of weights has identical values for the smallest feasible γ_κ value. For $\kappa = 1/N$ all weights are identical but for $\kappa = 0$ at least one weight is different from κ, but it also can be multiple weights. This is different if prior weights are used that are a feasible solution to the forecast combination problem. $\boldsymbol{\omega}^{IL}$ is a feasible solution, all weights ω_i are identical to the prior weight $\dot{\omega}_i$ for $\gamma_{\boldsymbol{\omega}^{IL}} = 0$, as depicted in Figure 4.9. Then, as $\gamma_{\boldsymbol{\omega}^{IL}}$ increases, single weights deviate from their prior weight (for example red and yellow), while others remain at theirs (teal, green, purple, and blue). As $\gamma_{\boldsymbol{\omega}^{IL}}$ increases further, those also deviate from their prior solution. For a sufficiently larger value of γ, i.e., $\gamma^*_{\boldsymbol{\omega}^{IL}}$, the solution is, again, identical to the optimal weights solution.

Note that the optimization problem of Equations (4.35) and (4.36) is a more general version of the problem of Equations (4.28) and (4.32), which used a fixed value of κ. Accordingly, we can derive the feasible interval of γ values similarly. The idea to derive

the smallest feasible value $\gamma_{\dot{\boldsymbol{\omega}}}$ is similar as for γ_κ from Section 4.2.2.1. We assume that $\omega_i = \dot{\omega}_i$, determine the sum of weights and calculate the difference to one. We need to allow for at least that much deviation from $\dot{\boldsymbol{\omega}}$ to ensure that weights can sum to unity. The feasible interval for $\gamma_{\dot{\boldsymbol{\omega}}}$ can be straightforward adapted from the smallest feasible value for a fixed value κ, see again Equation (4.29). To this end, we need to calculate the deviation of the sum of prior weights to one and take its absolute value. It corresponds to the smallest deviation that has to be allowed in order to fulfill the unity constraint, i.e.,

$$\gamma_{\dot{\boldsymbol{\omega}}} \in \left[\left|1 - \sum_{i=1}^{N} \dot{\omega}_i\right|, \infty\right). \tag{4.37}$$

If the prior weights vector $\dot{\boldsymbol{\omega}}$ is a feasible solution to the forecast combination problem, i.e., weights sum up to unity, the smallest feasible value for $\gamma_{\dot{\boldsymbol{\omega}}}$ is zero. Otherwise, it is greater zero to allow for the deviation of weights to the closest feasible solution. Equation (4.38) can be applied to any L_1 constraint forecast combination problem of our unified framework. To this end one has to define $\dot{\boldsymbol{\omega}} = (\kappa, \kappa, \ldots, \kappa)'$.

Similarly to the smallest feasible value, with prior weights we can provide a general formula to calculate γ^* for any of the considered L_1 methods, i.e., the largest value of γ that still constraints the solution space. The general formula is given as

$$\gamma^* = \|\boldsymbol{\omega}^{OW} - \dot{\boldsymbol{\omega}}\|, \tag{4.38}$$

i.e., the sum of absolute deviations of the OW solution and prior weights. If this amount of deviation is allowed, each weight will be equal to the optimal weight. In contrast, if we constrain the weights vector $\gamma < \gamma^*$, there is not enough possible deviation from prior weights such that the resulting weights can be equal to the optimal weights.

Beside the fact that shrinkage towards prior weights is more flexible and nests shrinkage towards fixed values, using prior weights enables even more possibilities for shrinkage directions. We will explore some additional shrinkage directions in the following section.

4.2.3 Conditional Group Equal Weights

Inspired by the LHS of Equation (4.2) and the peLasso of Equation (4.13) with $q = 1$, we define a new general approach to design prior weight vectors. It is build around the idea of the peLasso and how to simplify it into a one-step procedure while still using a quadratic optimization problem with linear constraints, see again Section 4.1.2. To this end, we developed two variants. Recall that the peLasso selects and shrinks towards

both towards zero and equal weights of the non-zero weight forecasts, i.e., $1/\|\boldsymbol{\omega}\|_0$. Within the forecast combination framework it is given by

$$\begin{aligned}
\underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} && \text{(4.39a)}\\
\text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, && \text{(4.39b)}\\
& \|\boldsymbol{w}\| \leq \gamma_0, && \text{(4.39c)}\\
& \|\boldsymbol{w} - 1/\|\boldsymbol{\omega}\|_0\| \leq \gamma_{1/\|\boldsymbol{\omega}\|_0} && \text{(4.39d)}
\end{aligned}$$

For the first simplification, we remove the fraction that includes the L_0 norm in Equation (4.13) and replace it by κ, i.e.,

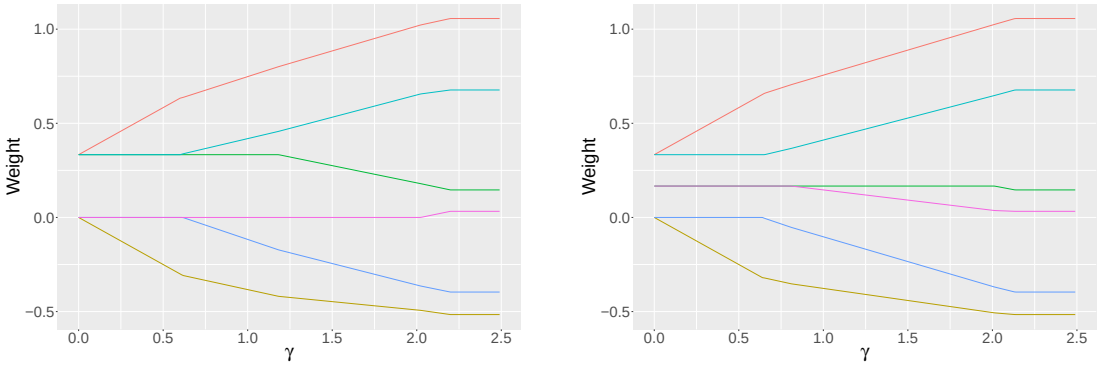
$$\begin{aligned}
\underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} && \text{(4.40a)}\\
\text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, && \text{(4.40b)}\\
& \|\boldsymbol{w}\| \leq \gamma_0, && \text{(4.40c)}\\
& \|\boldsymbol{w} - 1/\kappa\| \leq \gamma_\kappa && \text{(4.40d)}
\end{aligned}$$

The difference to the peLasso is in Equations (4.39d) and (4.40d). The optimization problem of Equation (4.40) simultaneously selects and shrinks a subset of weights to zero and κ. Accordingly, this method can select the subsets, however only for a given, arbitrary shrinkage direction. For example, shrinking a subset of weights of unknown size towards the equal weights of all forecast ($1/N$) is less intuitive compared to the equal weights of the number of non-zero weights for a given solution ($1/\|\boldsymbol{\omega}\|_0$).

The first variant we introduced can select which weights to shrink and select towards zero and which towards a fixed value κ. The second variant we propose uses a fixed subset of weights that it shrinks and selects towards zero while it shrinks and selects the other weights towards the equal weights of the latter subset. The prior selection of forecasts based on a given criterion, e.g., the forecast accuracy, is similar to the approach of LHS. Recall that LHS of Equation (4.2) also shrinks a predefined subset of weights towards zero and all other weights to their corresponding equal weights linearly starting from the optimal weight solution. By using the L_1 constraint, however, weights are also shrunken and selected towards either zero or the corresponding equal weights but not linearly as for the LHS.

Considering both variants we assess the second to be more sensible, because it can shrink towards the equal weights of a given subset instead of an arbitrary value κ without knowing the subset size. Moreover, as will show that we can extend the idea of the second variant which leads to a whole new approach to define a shrinkage direction using prior weights, i.e., we use the optimization problem of Equation (4.35).

To incorporate the simplified variant of the peLasso assume that we split forecasts into groups a priori. Now, we define the prior weights to be one of two possible values



(a) $G = 2$ with $g_1 = 1$ and $g_2 = 0$ **(b)** $G = 3$ with $g_1 = 2/3, g_2 = 1/3$ and $g_3 = 0$

Figure 4.10. Illustration of forecast weight paths new shrinkage directions with prior weights. Prior weights are determined by Equation (4.41) with two and three groups. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

depending on which group a forecast belongs to: either zero or the equal weights of all forecast within the corresponding group. Again, note that this is basically the shrinkage direction of the LHS proposed in Schulz et al. (2022). However, we shrink towards it not linearly but using a L_1 constraint and also select weights to be exactly equal to the shrinkage direction. By that we combined the ideas from the peLasso, in a simplified way, and the LHS.

Moreover, we adapt the approach and propose a new generalized shrinkage direction. We refer to it as shrinkage towards *Conditional Group Equal Weights (CGEW)*. To this end, we divide the N forecasts into G groups or subsets, i.e., $\mathbb{G}_j \ \forall j = 1, \ldots, G$. Each group $\mathbb{G}_j$ is assigned a proportion or budget $g_j \in [0, 1]$. To put it differently, it defines how much of the solution for the smallest feasible γ value each group gets. Accordingly, it has to hold that $\sum_{j=1}^{G} g_j = 1$. The budget is then distributed equally among all forecasts within a group.[25] By that, we define the prior weight of forecast i where $i \in \mathbb{G}_j$ to be the *conditional equal weights* of the corresponding group based on the assigned budget, i.e.,

$$\dot{\omega}_i = \frac{g_j}{|\mathbb{G}_j|} \qquad \forall i \in \{1, \ldots, N : i \in \mathbb{G}_j, j = 1, \ldots, G\}. \tag{4.41}$$

As a result, all weights or forecasts i within the same group $\mathbb{G}_j$ are shrunken and selected towards this conditional equal weights of the group that is defined by the assigned budget.

[25] If N/G is odd, we assign $\lfloor N/G \rfloor$ forecasts to each group $j = 1, \ldots, G-1$ and $N - \sum_{j=1}^{G-1} |\mathbb{G}_j|$ forecasts to group G. To put it differently, the last group gets an additional forecast.

Figure 4.10 depicts the weight paths of two variants of the new shrinkage direction that we will use within this thesis. First, we use two groups ($G = 2$) where the 50% of forecasts with the best forecast accuracy are within group one and all others are within group two. The first group gets the whole available budget, i.e., $g_1 = 1$. As a consequence, forecasts within group one are shrunken towards their equal weights and forecasts from group two towards zero, see again Equation (4.41). We will refer to these prior weights by $\boldsymbol{\omega}^{E2}$ and it is basically the shrinkage direction of the LHS method. The weight path of it is depicted in Figure 4.10(a). The best three forecasts (red, teal, and green) are shrunken and selected from the optimal weights solution for about $\gamma_{\boldsymbol{\omega}^{E2}} \geq 2.5$ towards their equal weights, i.e., $1/3$, as $\gamma_{\boldsymbol{\omega}^{E2}}$ approaches zero. In contrast, the three forecasts with the highest MSE (yellow, blue, and pink) are shrunken and selected towards zero.

The second variant that we consider uses three groups ($G = 3$) and is denoted by $\boldsymbol{\omega}^{E3}$. The first group gets $g_1 = 2/3$ of the budget and the second group $g_2 = 1/3$ leaving no budget for group three ($g_3 = 0$). The weight paths are depicted in Figure 4.10(b). The best two forecasts (red and teal) from group $j = 1$ are shrunk towards their conditional equal weights based on the given budget of $\frac{2/3}{2} = 1/3$, see Equation (4.41). The next two forecasts (green and pink) of group $j = 2$ with budget $g_2 = 1/3$ are shrunken towards their prior weights of $1/6$. The two forecast from group $j = 3$ with the largest MSE in the training set (blue and yellow) are shrunken towards their prior weights of zero. The corresponding weight paths are depicted in Figure 4.10(b). Each group is shrunken towards their prior weights as $\gamma \to 0$. For example, given $\gamma_{\boldsymbol{\omega}^{E3}} = 0.5$, only the red and yellow forecast deviate from their conditional equal weights.

Within this thesis, we consider shrinkage direction for CGEW, were forecasts with higher accuracy are shrunk and selected to higher weights.

In summary, inspired by mainly LHS and peLasso we propose to use CGEW as a shrinkage directions with the L_1 norm. To this end, we sort forecast into groups based on their forecast accuracy and then shrink each group of forecasts towards the conditional equal weights based on a predefined budget. We can also interpret this as a sophisticated and comprehensive hybrid between $L_1(0)$ and $L_1(1/N)$. The introduction of this shrinkage direction to be used with a L_1 constraint is the third main contribution of this thesis.

4.2.4 Summary

In summary, we presented a unified framework that incorporates the L_1 constraint into the forecast combination problem as originally defined in J. M. Bates and Granger (1969). To this end, we use prior weights as a shrinkage direction, see Equations (4.35) and (4.36). Our unified framework nests a Lasso with a unity constraint as considered by Radchenko et al. (2023), i.e., shrinkage towards zero. Furthermore, it incorporates the

egalitarian Lasso (shrinkage towards equal weights) with a unity constraint proposed by Diebold and Shin (2019) and enables shrinkage towards a prior weights vector as done in Roccazzella et al. (2022). Moreover, our optimization problem of Equation (4.36) is not limited to these cases. It can be used to shrink and select weights towards any fixed value κ or any prior weights vector $\dot{\omega}$. It provides a unified framework that is based on the OW forecast combination problem with a unity constraint and an objective function to minimizes the combined error variance proposed by J. M. Bates and Granger (1969). Additionally, it can be used to define the partially egalitarian Lasso in a one-step procedure, which was previously left for future research by Diebold and Shin (2019). However, the resulting optimization problem is a quadratic mixed integer problem, see Appendix A. Lastly, we proposed a to use a shrinkage direction based on prior weights. We refer to it as shrinkage towards conditional group equal weights (CGEW). To this end, forecasts are sorted into groups based on their forecast accuracy and those groups are shrunken towards a conditional equal weights of the group defined by the groups budget.

As the unified framework in the form of the quadratic optimization problem with prior weights allows for a direct comparison on the same basis between the different L_1 norm approaches, we will analyze those approaches and additional benchmarks with respect to their forecast accuracy.

4.3 Application: Simulation Study

In this section we analyze and compare forecast accuracy of the forecast combination methods from Section 4.2. We use the simulation study introduced in Section 3.2. For a brief summary of the designed scenarios see Section 3.2.3 and the used forecast error correlation matrices are depicted in Section 3.2.2.2.

Both the L_1 and LHS methods, have at least one hyperparameter (γ and λ). In real world applications, those hyperparameters have to be estimated if the method is used to forecast future values. However, we follow related literature (see e.g., Diebold & Shin, 2019; Radchenko et al., 2023; Roccazzella et al., 2022) and split the analysis into two parts. In the first part, Section 4.3.1, we analyze the forecast accuracy without hyperparameter estimation. By that, we can compare the combination methods without the uncertainty introduced by hyperparameter estimation. Accordingly, we *ex post* choose the best hyperparameter values for each method. For the second part of the analysis, Section 4.3.2, we analyze the capabilities of the forecast combination methods to be used in an actual forecast setting. To this end, we use cross-validation to estimate the hyperparameters, compute out-of-sample forecasts and assess the performance of the forecast combination methods. Lastly, we summarize our findings for both the ex post and pseudo out-of-sample analysis in Section 4.3.3. Although, the ex post analysis

we also compute out-of-sample forecasts, we decided to differentiate the two analysis by the terms: *ex post* and *out-of-sample* analysis. The terms basically refer to how the hyperparameters are determined. Either retrospectively, i.e., ex post, or they have to be estimated based only on past information and thus the true out-of-sample forecast accuracy is evaluated.

Regardless of which analysis is performed, first, we need to choose a grid of candidate values for the hyperparameters.

Grid of Candidate Values The grid of candidate values for LHS is straightforward by its design. The method shrinks weights from the optimal weight solution ($\lambda = 0$) to either equal weights and/or zero ($\lambda = 1$). Thus, we use a grid from zero to one with 0.1 increments. For the L_1 methods, the grid needs to be chosen more carefully. First, for a given set of observations, recall that we can determine the largest value of the hyperparameter, i.e., γ^*. It is the sum of the absolute difference between the optimal weights and prior weights $\dot{\boldsymbol{\omega}}$, see Equation (4.38). Accordingly, we can determine which γ-values we need to consider. However, even for another ever so slightly different set of observations, γ^* can be different. However, for our analysis for each observation in the test set we need forecasts for the same hyperparameter values. If forecasts are missing for some hyperparameter values, we can not properly evaluate the forecast accuracy. To this end, we choose a very large value as the end of the search grid. Then for each set of observations we only calculate forecasts up to the current γ^* and assign the forecast for γ^* to all candidate values $\gamma > \gamma^*$.

This is particularly well applicable if only a few or short time series are given. However, within a simulation study, an extensive empirical analysis or in a business context where one needs forecasts for a multitude of time series regularly, it becomes more challenging due to limited computational resources. For example, the evaluation of all potential values takes too long or statistical software can get slow because the resulting objects with weights and forecasts become too large. In this case one can either limit the candidate values due to preference or prior information, or increase the increment between candidate values.

In this thesis, we use a grid that starts at the smallest feasible value of γ of each forecast combination method and ends at 50 with 0.1 increments. Note that we defined this grid prior to the analysis. An inspection of the ex post results afterwards revealed that, for example, for both $L_1(0)$ and $L_1(1/N)$ the largest ever observed γ^* over all test sets and scenarios was below 36.

4.3.1 Ex Post Analysis: L_1 and LHS

In the simulation study we use 200 time series with 90 observations each. The last 50 observations of each time series are used as a test set. Accordingly, we have a total of $10,000$ observations that are used to evaluate the forecast accuracy of the forecast combination methods. For each time series we perform pseudo out-of-sample forecasting with a rolling window of size 40, see again Section 2.1. If a forecast combination method does not have any hyperparameters like EW, PW, OW or IL we can simply calculate forecasts for each observation of the test set. In contrast, for each method that has hyperparameters, i.e., LHS, $L_1(\kappa)$, and $L_1(\dot{\boldsymbol{\omega}})$, we use a grid of candidate values for each hyperparameter. For all candidate values we then estimate weights and compute forecasts. Afterwards (ex post) we choose from the hyperparameter values that result in the smallest MSE over the whole test set. To put it differently, if we use this fixed value for every observation within this test set, it results in the smallest MSE compared to all other candidate values. We repeat this process for all 200 time series and average the individual 200 MSE values. For the sake of clarity, we chose the best hyperparameter for each test set individually. By that, hyperparameters can be chosen specifically tailored towards the current development of the times series.

In what follows we, first, analyze the results for all scenarios in Section 4.3.1.1 together. Second, in Section 4.3.1.2 we analyze the results with respect to the error correlation matrices, error variance similarities and special groups, i.e., groups of scenarios defined within our simulation study.

4.3.1.1 Ex Post Analysis: Overall Results

To analyze the results of the simulation study, we will use a combination of tables, figures and summary statistics. Tables 4.1 and 4.2 present the MSE values of the different methods for all considered scenarios. Each row represents a scenario that is defined by the first three columns that consist of the correlation matrix, CM, the error variance similarity, z, and special groups, SG. Table 4.1 presents the results for CM1 to CM3 and Table 4.2 for CM4 to CM6. The forecast combination methods are depicted in the remaining columns. The results of the benchmark methods EW, PW, OW and IW are reported in the columns four to seven. Columns eight and nine show the results for the LHS methods if either all forecasts are used (LS) or if a subset (LHS) is used.[26] The L_1 constraint methods that shrink towards a fixed value of zero or $1/N$ are shown in columns ten and eleven. The last three columns show the results for the L_1 constraint with shrinkage towards a prior weights vector. Columns twelve and thirteen depict the results of the prior weights vectors we proposed to use in Section 4.2.3, i.e., shrinkage towards conditional group equal weights (CGEW). The last column presents the results

[26] Note that, for the sake of simplicity, if we refer to both LS and LHS we will denote it by LHS methods, because LS is a special case of LHS.

for weights shrunken towards the inverse-loss weights of Equation (4.16). Henceforth, we will refer to scenarios by CM/z/SG. For example, consider scenario 1/0.05/none, i.e., the first row in Table 4.1. The EW, IL and PW methods have similar MSE values with 0.98, 0.98 and 0.97. The OW approach has a MSE about twice as high with 2.11. The LHS methods can reduce the MSE by 0.01 and 0.02 in comparison to the best benchmark method, PW. If an L_1 constraint is used the MSE is even smaller except $L_1(0)$. The smallest MSE of 0.92 is given for the shrinkage towards prior weights $\boldsymbol{\omega}^{EW2}$ and $\boldsymbol{\omega}^{EW3}$ (CGEW).

In our analysis, we have eleven different methods and 72 scenarios. This renders an analysis with the level of detail shown exemplary above to be impractical or impossible. Especially, if we not only consider single scenarios but draw conclusion for overarching properties between partially common scenarios, e.g., compare method with respect to error correlation matrices. For comparison, recall that Roccazzella et al. (2022) also considers the Lasso based method with a Lagrangian relaxation, and we use their simulation study as a baseline from we build our simulation study. In their analysis they use four scenarios while we are using 72. Furthermore, we both consider an ex post and pseudo out-of-sample analysis.[27] Therefore, we analyze and summarize the results of Tables 4.1 and 4.2 to draw conclusions.

Table 4.3 depicts three different metrics for all methods (columns). The first row presents the percentages of how often each method has the smallest MSE over all scenarios. Note that the percentages add up to a value greater one, because the average MSE over all test sets is rounded to the second decimal place and, as a result, there can be multiple methods with the same smallest MSE. Henceforth, we will use *smallest MSE* as a more lose description, i.e., there can be multiple method that have the same value. If we want to emphasize that a method has the smallest method overall, we use *strictly smallest MSE*.

The second row shows the average rank over all scenarios. For the average rank, we rank the methods for each scenario. The method with the strictly smallest MSE gets rank 1, the method with the second smallest MSE rank 2 and so on. If multiple methods have the same MSE, they are temporarily ranked in an arbitrary order and then the average rank is computed and used for each method.[28]

The third row depicts the average distance or difference in MSE to the scenario-wise best method. For the average distance we first calculate the difference between the MSE of each method and the smallest MSE within each scenario. Then the distance is averaged across scenarios. The average rank and distance provide additional information besides the percentages of how often one method has the smallest MSE. With the rank

[27] Note that neither Radchenko et al. (2023) nor Diebold and Shin (2019) analyze the out-of-sample forecast accuracy of the lasso-based methods in terms of a simulation study.

[28] With respect to the average ranks we follow the approach of the Spearman correlation coefficient, see Fahrmeir et al. (2016, pp. 133-135).

CM	z	SG	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\omega})$		
							LS	LHS	0	$1/N$	ω^{E2}	ω^{E3}	ω^{IL}
1	0.05	none	0.98	0.97	2.11	0.98	0.96	0.95	0.95	0.93	**0.92**	**0.92**	0.93
		first	0.97	0.93	1.89	0.96	0.93	0.92	0.89	0.89	**0.88**	**0.88**	0.89
		last	1.01	0.99	2.04	1.00	0.98	0.96	0.95	0.94	**0.93**	**0.93**	0.94
		both	0.97	0.93	1.75	0.96	0.91	0.89	0.87	0.87	**0.85**	**0.85**	0.86
1	0.20	none	1.16	0.92	1.03	1.12	0.76	0.73	0.60	0.60	**0.59**	**0.59**	0.60
		first	1.12	0.73	0.52	1.04	0.44	0.42	**0.30**	0.31	0.31	0.31	0.31
		last	1.24	0.95	0.84	1.18	0.66	0.63	0.48	0.48	**0.47**	**0.47**	0.48
		both	1.15	0.72	0.45	1.05	0.39	0.38	**0.26**	0.27	**0.26**	**0.26**	0.27
1	0.50	none	1.53	0.80	0.42	1.33	0.39	0.37	**0.24**	0.25	**0.24**	**0.24**	0.25
		first	1.38	0.30	0.09	0.85	0.09	0.09	**0.05**	0.06	**0.05**	**0.05**	**0.05**
		last	1.64	0.82	0.35	1.38	0.33	0.32	**0.20**	0.21	**0.20**	**0.20**	**0.20**
		both	1.46	0.30	0.08	0.85	0.08	0.08	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
2	0.05	none	0.56	0.62	1.37	0.56	0.56	0.57	0.62	**0.55**	0.57	0.56	**0.55**
		first	0.56	0.62	1.34	0.56	0.56	0.57	0.62	**0.55**	0.57	0.56	**0.55**
		last	0.57	0.64	1.37	0.57	0.57	0.58	0.63	**0.56**	0.58	0.57	**0.56**
		both	0.57	0.63	1.35	0.57	0.57	0.58	0.63	**0.56**	0.57	0.57	**0.56**
2	0.20	none	0.68	0.68	1.42	0.66	0.66	0.64	0.67	0.64	**0.62**	**0.62**	0.63
		first	0.64	0.56	1.11	0.60	0.60	0.55	0.54	0.56	**0.53**	**0.53**	0.54
		last	0.68	0.66	1.31	0.65	0.66	0.61	0.65	0.62	**0.58**	0.59	0.61
		both	0.66	0.54	1.00	0.60	0.57	0.52	0.51	0.53	**0.50**	**0.50**	0.51
2	0.50	none	0.87	0.64	1.10	0.76	0.69	0.62	0.58	0.61	**0.56**	0.57	0.59
		first	0.80	0.26	0.35	0.49	0.30	0.28	**0.20**	0.23	0.22	0.22	0.21
		last	0.96	0.66	1.05	0.81	0.71	0.62	0.57	0.60	**0.56**	**0.56**	0.58
		both	0.83	0.26	0.33	0.49	0.29	0.27	**0.19**	0.22	0.21	0.21	0.20
3	0.05	none	**0.25**	0.32	0.63	0.26	**0.25**	0.29	0.32	**0.25**	0.28	0.27	**0.25**
		first	0.25	0.31	0.63	0.25	0.25	0.28	0.31	**0.24**	0.27	0.26	**0.24**
		last	**0.25**	0.32	0.63	**0.25**	**0.25**	0.28	0.32	**0.25**	0.28	0.27	**0.25**
		both	0.25	0.32	0.64	0.25	0.25	0.28	0.32	**0.24**	0.27	0.26	0.25
3	0.20	none	0.30	0.36	0.71	0.30	0.30	0.31	0.36	**0.29**	0.31	0.30	**0.29**
		first	0.29	0.32	0.63	**0.27**	0.29	0.28	0.32	0.28	0.28	**0.27**	**0.27**
		last	0.30	0.36	0.71	**0.29**	0.30	0.31	0.36	**0.29**	0.30	**0.29**	**0.29**
		both	0.30	0.32	0.61	0.27	0.29	0.28	0.32	0.28	0.27	**0.26**	0.27
3	0.50	none	0.40	0.39	0.75	0.35	0.38	0.35	0.39	0.37	0.34	**0.33**	0.34
		first	0.35	**0.19**	0.35	0.22	0.24	0.22	**0.19**	0.22	0.21	0.21	**0.19**
		last	0.43	0.40	0.76	0.36	0.40	0.35	0.40	0.38	0.34	**0.33**	0.35
		both	0.37	0.20	0.36	0.22	0.25	0.22	**0.19**	0.23	0.21	0.21	**0.19**

Table 4.1. Simulation study results of benchmark, LHS and L_1 methods for correlation matrices CM1, CM2 and CM3 (ex post analysis). The table depicts the MSE of the forecast combination method of Section 4.2 and other benchmarks. The methods with the smallest MSE are depicted in bold numbers.

we can determine how consistent the forecasting performance a method is compared to others. For example, assume that there are ten methods with unique MSE values within each scenario. Among them there is method A which has the best forecast accuracy for half of the scenarios (rank 1) and the worst forecast accuracy for the other half (rank 10). Let there be another method B that never has the best forecast accuracy but has the second best (rank 2) consistently throughout all scenarios. As a consequence, the average rank of method A (5.5) is higher than that of method B (2). To make an informed decision which method to apply to a real-world applications, a more consistent method can be preferable. Especially, if there are no further information as,

CM	z	SG	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\omega})$		
							LS	LHS	0	$1/N$	ω^{E2}	ω^{E3}	ω^{IL}
4	0.05	none	**0.63**	0.69	1.57	**0.63**	**0.63**	0.69	0.68	**0.63**	0.67	0.66	**0.63**
		first	0.65	0.70	1.61	0.65	0.65	0.70	0.70	**0.64**	0.68	0.67	**0.64**
		last	0.66	0.72	1.64	0.66	0.66	0.72	0.72	**0.65**	0.70	0.69	**0.65**
		both	0.64	0.68	1.59	0.63	0.64	0.68	0.68	**0.62**	0.66	0.65	0.63
4	0.20	none	0.79	0.78	1.78	0.76	0.78	0.77	0.78	**0.74**	0.76	0.75	**0.74**
		first	0.73	0.67	1.48	0.68	0.71	0.67	0.66	0.65	0.65	**0.64**	**0.64**
		last	0.80	0.78	1.74	0.77	0.79	0.76	0.77	**0.74**	0.75	**0.74**	**0.74**
		both	0.77	0.68	1.51	0.70	0.74	0.68	0.67	0.66	0.66	**0.65**	0.66
4	0.50	none	1.03	0.79	1.50	0.91	0.88	0.81	**0.75**	0.77	0.77	0.76	0.77
		first	0.90	0.30	0.45	0.56	0.37	0.35	**0.24**	0.26	0.26	0.26	0.25
		last	1.10	0.79	1.47	0.93	0.91	0.81	**0.75**	0.77	0.77	0.77	0.77
		both	1.02	0.31	0.46	0.60	0.39	0.36	**0.25**	0.26	0.26	0.26	0.26
5	0.05	none	0.44	0.40	0.84	0.45	0.42	0.48	0.40	**0.38**	0.41	0.41	0.39
		first	0.44	0.40	0.85	0.45	0.42	0.49	0.40	**0.38**	0.41	0.41	0.39
		last	0.44	0.41	0.88	0.46	0.43	0.50	0.40	**0.38**	0.41	0.41	0.39
		both	0.43	0.41	0.87	0.45	0.42	0.49	0.41	**0.38**	0.41	0.41	0.39
5	0.20	none	0.51	0.51	1.06	0.54	0.50	0.59	0.51	**0.46**	0.51	0.50	0.47
		first	0.48	0.49	0.98	0.50	0.47	0.54	0.48	**0.44**	0.49	0.48	0.45
		last	0.52	0.54	1.13	0.56	0.51	0.62	0.54	**0.49**	0.54	0.54	0.51
		both	0.49	0.53	1.06	0.51	0.48	0.56	0.52	**0.47**	0.51	0.50	0.48
5	0.50	none	0.67	0.66	1.21	0.69	0.63	0.70	0.65	**0.61**	0.65	0.64	0.62
		first	0.56	0.31	0.42	0.45	0.32	0.32	**0.27**	0.29	0.30	0.30	0.29
		last	0.66	0.67	1.22	0.68	**0.62**	0.68	0.66	**0.62**	0.65	0.65	**0.62**
		both	0.59	0.33	0.44	0.46	0.34	0.34	**0.29**	0.31	0.32	0.31	0.31
6	0.05	none	0.79	0.63	1.39	0.77	0.73	0.72	0.62	**0.60**	0.61	0.61	0.61
		first	0.78	0.60	1.32	0.75	0.72	0.69	0.59	**0.58**	**0.58**	**0.58**	**0.58**
		last	0.80	0.63	1.37	0.78	0.74	0.72	0.62	**0.61**	**0.61**	**0.61**	**0.61**
		both	0.80	0.61	1.30	0.77	0.72	0.69	0.60	**0.58**	**0.58**	**0.58**	0.59
6	0.20	none	0.97	0.64	1.16	0.89	0.75	0.67	0.59	0.58	**0.56**	**0.56**	0.58
		first	0.94	0.52	0.95	0.81	0.66	0.57	0.48	0.48	**0.45**	**0.45**	0.47
		last	1.02	0.64	0.96	0.91	0.68	0.61	0.53	0.52	**0.50**	**0.50**	0.52
		both	0.97	0.51	0.80	0.81	0.59	0.52	0.44	0.43	**0.40**	0.41	0.42
6	0.50	none	1.28	0.61	0.85	1.02	0.67	0.58	0.48	0.47	**0.44**	0.45	0.47
		first	1.22	0.26	0.42	0.62	0.37	0.33	0.21	0.22	**0.20**	0.21	0.21
		last	1.39	0.62	0.69	1.05	0.57	0.51	0.40	0.39	**0.37**	**0.37**	0.38
		both	1.31	0.26	0.38	0.62	0.34	0.31	0.20	0.20	**0.19**	**0.19**	0.20

Table 4.2. Simulation study results of benchmark, LHS and L_1 methods for correlation matrices CM4, CM5 and CM6 (ex post analysis). The table depicts the MSE of the forecast combination method of Section 4.2 and other benchmarks. The methods with the smallest MSE are depicted in bold numbers.

for example, method A is always good for highly correlated forecast errors. The average distance is useful to be able to further assess the consistency and overall performance of a method. Methods with a similar average rank nevertheless can have very different average distances. In such a case the method that has, on average, a smaller difference in MSE to the best method is preferable.

We use the information of Table 4.3 in addition to Figure 4.11. The latter figure provides an overview of the ranks for each methods over all scenarios. The box depicts the middle 50%, i.e., it ranges from the 75% quantile to the 25% quantile. This range is called *interquartile range (IQR)*. The line within the box is the median. The lines

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\omega})$		
					LS	LHS	0	$1/N$	ω^{E2}	ω^{E3}	ω^{IL}
Smallest MSE (%)	4.17	1.39	0.00	5.56	5.56	0.00	22.22	**43.06**	38.89	**43.06**	33.33
Avg Rank	8.60	7.42	10.31	7.69	6.59	7.10	5.24	3.26	3.83	3.20	**2.74**
Avg Distance	0.31	0.10	0.54	0.21	0.08	0.07	0.03	**0.01**	0.02	**0.01**	**0.01**

Table 4.3. Key figures for the MSE values of benchmark, LHS and L_1 methods over all simulation study scenarios (ex post analysis). Smallest MSE (%) - Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank - Average rank of a method where a smaller rank is favorable. Avg Distance - Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

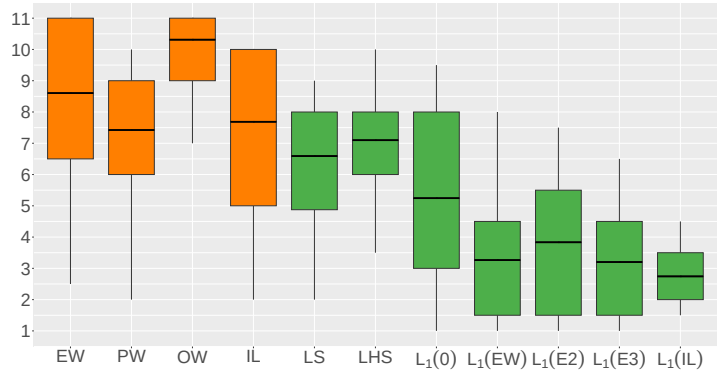


Figure 4.11. Boxplot of ranks across benchmarks, LHS and L_1 methods for the ex post analysis.

outside the boxes are called whiskers and are 1.5 times the length of the IQR or up to the largest or smallest observation respectively. The colors in Figure 4.11 distinguish between the benchmark and LHS / L_1 methods.

First, let us consider the benchmark methods using Figure 4.11. The middle 50% of OW range from about nine to eleven, i.e., for 75% of scenarios, OW has a rank of nine or higher. Table 4.3 shows that OW method is never the best method, and it has both the largest average rank and distance.

Note that a larger range of ranks can be both advantageous and disadvantageous. If the overall magnitude of ranks is small, a small range of ranks is better. However, if the overall magnitude of ranks is large, a wider range indicates that a method can be good in certain scenarios. For example, OW has overall large average ranks and a small range of ranks, i.e., usually it is among the worst method and there less or no scenarios where it is actually good. In comparison, PW has a smaller average rank but a slightly larger range of ranks. Nevertheless, as Figure 4.11 clearly shows, it is preferable as it achieves smaller ranks frequently. Both PW and IL have similar average ranks, see Table 4.3. However, the range of ranks is larger for IL. If we compare both the 25% and 75% of PW and IL, we can see that although IL has better ranks for some scenarios, the opposite

holds true for others. An example where a larger range of ranks is advantageous is LS and LHS. They have about the same 75% quantile, i.e., for 75% of the scenarios their rank is about eight. However, the 25% quantile of LS is roughly five, while for LHS it is about six. Accordingly, LS achieves smaller ranks more regularly compared to LHS. We will use this kind of comparison throughout this thesis.

The four benchmarks, LS and LHS have the smallest MSE for the least amount of scenarios, their average ranks and distances are the highest. With respect to the benchmark methods, PW is preferable because it has the smallest average rank, preferable rank distribution and smallest average distance to the best method. However, the same argumentation holds true if one compares PW to LS and LHS as one can immediately see in Figure 4.11. If we compare the two LHS methods, LS, i.e., shrinking all weights towards equal weights instead of shrinking some towards zero is slightly better (Best (%), average rank and rank distribution). Table 4.3 shows that there are scenarios where EW, PW, IL and, LS have the smallest MSE. However, the result from Tables 4.1 and 4.2 show that they never have the strictly smallest MSE. There are other method beside the benchmark methods that have the same MSE.

Let us now consider the methods that use a L_1 constraint to shrink weights towards either a fixed value $L_1(\kappa)$ or prior weights $L_1(\dot{\boldsymbol{\omega}})$. Both Table 4.3 and Figure 4.11 provide clear evidence that the L_1 methods generally improve the forecast accuracy within our simulation study. All of them have the smallest MSE more often, smaller average ranks and distance than the benchmarks and LHS methods. Moreover, the rank distribution is also beneficial, i.e., the middle 50% have smaller ranks in comparison.

$L_1(0)$ stands out noticeably from the L_1 methods as it has a smaller percentage of having the smallest MSE (22.22%). Additionally, its average rank and IQR is the largest among all L_1 methods and the middle 50% include the largest ranks. The other L_1 methods are the best method more often, particularly $L_1(1/N)$ and $L_1(\boldsymbol{\omega}^{E3})$ with 43.06% of scenarios. $L_1(\boldsymbol{\omega}^{E3})$ overall have the

Shrinkage towards prior weights $\boldsymbol{\omega}^{E2}$ or $\boldsymbol{\omega}^{E3}$ as well as $L_1(1/N)$ have a 25% quantile of about 1.5. i.e., for 18 out of 72 scenarios they have the smallest MSE.[29] Both $L_1(1/N)$ and $L_1(\boldsymbol{\omega}^{E3})$ have a similar average (3.26 and 3.20) and range of ranks. However, the smallest average rank (2.74) is achieved by $L_1(\boldsymbol{\omega}^{IL})$. $L_1(1/N)$ and $L_1(\boldsymbol{\omega}^{E3})$ reach smaller ranks for some scenarios, but also larger ranks for others. Accordingly, the smaller IQR of $L_1(\boldsymbol{\omega}^{IL})$ can be considered advantageous as it has small ranks more consistently over many scenarios.

In summary, L_1 methods enhance the forecast accuracy compared to both the benchmark and LHS methods. Shrinkage towards equal weights as well as other, prior weights is favorable compared to shrinking all weights towards zero.

[29] Recall that two methods can have the same MSE and are then the average rank is used.

In what follows we analyze the forecasting performances of the methods with respect to groups of scenarios, i.e., the correlation matrices CM, error variance similarities z and special groups SG.

4.3.1.2 Ex Post Analysis: Groups of Scenarios

Let us now take a closer look at the forecast accuracy with respect to the design of the different scenarios. To this end, we group the scenarios with respect to the correlation matrix CM, error variance similarities z or special groups SG and aggregate the results. Note that due to the large extend of the simulation study we focus our analysis on the more advanced, shrinkage based forecast combination methods. We will compare them between each other and benchmark methods. However, we will less focus on comparing different benchmark methods with each other.

The analysis is still based on the results presented within Tables 4.1 and 4.2. Moreover, we will also consider Smallest MSE (%), rank and distance similar to Table 4.3. While the percentage of how often a method has the smallest MSE is still presented in a table we use a different illustration for average ranks and differences. This is due to the fact that we have to differentiate between six correlation matrices, three z values and four special groups. Due to the large extent of this analysis, also consider the distribution of ranks for each group in addition to the average rank and distances is beyond the scope of this thesis.

Before we analyze the results, we briefly introduce the illustration of average rank and distance within Figures 4.12 and 4.13. The abscissa depicts the different methods in the same order as Tables 4.1 and 4.2.[30] In case of Figure 4.12 the ordinate shows the error correlation matrices. As a reference point, *all* presents the average ranks and differences of each method over all scenarios, i.e., the result of Section 4.3.1.1. In Figure 4.13 the ordinate shows, beside *all*, the different values for z and SG.

For illustration purposes, let us focus on Figure 4.12. For each method and error correlation matrix pair there is a colored circle. The size of the circle depicts the average rank. A larger circle indicates a better, i.e., smaller, average rank. The average distance between the MSE of a method and the best method within each scenario is depicted by color. It starts at yellow for a distance of zero, goes over green to blue and ends at purple (roughly 0.5). A legend for both the size of the circle and color is provided on the right side of the figure. Accordingly, a method with a large yellow circle indicates a small average rank and distance, i.e., it is preferable. Note that we trimmed the distances artificially. All distances that are more than two standard deviations away from the overall mean distance, get a black color. Otherwise, those larger values

[30] Note that we use a slightly different notation here due to the fact that there is no adequate depiction for the abbreviation of the methods in R (R Core Team, 2022).

distort the color gradient and differences in MSE distances between methods can not be identified anymore.

Table 4.4 and Figure 4.12 present the results of the forecast combination methods with respect to the error correlation matrices. The table shows the percentages of how often a method (columns) has, potentially among others, the smallest MSE with respect to the error correlation matrices (rows).

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
CM1	0.00	0.00	0.00	0.00	0.00	0.00	50.00	8.33	**91.67**	**91.67**	25.00
CM2	0.00	0.00	0.00	0.00	0.00	0.00	16.67	33.33	**50.00**	33.33	33.33
CM3	16.67	8.33	0.00	25.00	16.67	0.00	16.67	50.00	0.00	41.67	**66.67**
CM4	8.33	0.00	0.00	8.33	8.33	0.00	33.33	**50.00**	0.00	25.00	**50.00**
CM5	0.00	0.00	0.00	0.00	8.33	0.00	16.67	**83.33**	0.00	0.00	8.33
CM6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.33	**91.67**	66.67	16.67

Table 4.4. Percentage of scenarios for which benchmarks, LHS and L_1 methods have the smallest MSE with respect to the error correlation matrix (ex post analysis). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

The benchmarks methods have the smallest MSE for only a few scenarios for CM3 and CM4. The benchmark methods have the smallest MSE for some scenarios in CM3 and one in CM4. For the EW this is most likely due to the fact that CM3 (low error correlation of 0.2) is closer to the situation for which equal weights is optimal, i.e., uncorrelated forecasts errors and identical error variances recall Section 2.2.3. LS also has the smallest MSE in CM3 and CM4 for some scenarios but for one scenario of CM5. In contrast, LHS has never the smallest MSE. In comparison, the L_1 methods as a whole have the smallest MSE noticeably more often for all error correlation matrices.

The superiority of the L_1 methods is supported by Figure 4.12. Overall they have smaller average ranks (larger circle) and distances (yellow color). The original forecast combination methods without any constraints (OW) has the worst forecast accuracy of all (large average ranks and distances). The closest any benchmark and LHS method gets to the L_1 methods is for CM3. However, it is not EW as one may expect for a low error correlation but IL.

For highly correlated forecast errors within CM1 (0.9 error correlation), the shrinkage directions we propose in the context of the L_1 constraint, $L_1(\dot{\boldsymbol{\omega}})$ with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ have the smallest MSE for all scenarios but one (1/0.2/first), see Table 4.4. This hold accordingly in case of $\boldsymbol{\omega}^{E2}$ and CM6. This error correlation matrix has overall high correlations (0.9) except for the group with the best forecast accuracy (medium correlation 0.5). The best average ranks for both CM1 and CM6 are achieved by $\boldsymbol{\omega}^{E2}$

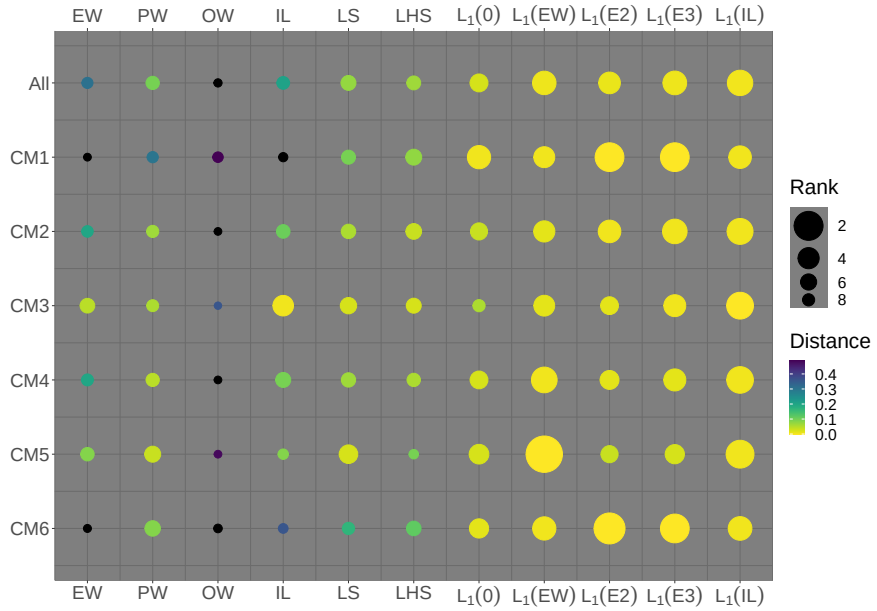


Figure 4.12. Illustration of average ranks and distances of the benchmark, LHS and L_1 methods for different correlations matrices (ex post analysis).

and $\boldsymbol{\omega}^{E3}$, see Figure 4.12. Accordingly, for highly correlated forecast errors shrinking some weights to zero while others to a conditional equal weights using an L_1 constraint is superior to its related methods. Those are LHS which also has can have the group structure but shrinks weights linearly, or $L_1(\kappa)$ which shrinks all weights to either zero or EW.

As the error correlation becomes smaller, i.e., as we move from CM1 to CM3 $L_1(\dot{\boldsymbol{\omega}})$ with $\boldsymbol{\omega}^{IL}$ becomes the best method more often and has a better average rank. This is somewhat surprising, because smaller error correlations tend towards equal weights being optimal (no error correlation between forecasts). However, the different considered error variance similarities (0.05, 0.2 and 0.5) lead to a decrease in performance of, e.g., $L_1(1/N)$. That in conjunction with the small deviation from no error correlation between forecasts appears to be sufficient that shrinkage towards $\boldsymbol{\omega}^{IL}$ is favorable over all EW, shrinkage towards EW with LHS or L_1 as well as shrinkage towards zero and conditional equal weight ($\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$) with an L_1 constraint.

For CM4 groups of forecast are highly correlated while the error correlation between groups is medium. Both, $L_1(1/N)$ and $L_1(\boldsymbol{\omega}^{IL})$ have the smallest MSE for about the same six out of twelve scenarios. This is reflected by the average rank and distance which are close to each other, see Figure 4.12. Although, the average rank of L_1 is noticeably smaller, it has the strictly smallest MSE for all other scenarios with the highest error variance similarity ($CM4/0.5/\cdot$).

For CM5 the better a group of forecasts is, the higher the error correlation is. Interestingly, $L_1(1/N)$ has the (strictly) smallest MSE for (eight) ten of the twelve scenar-

ios. Although,$L_1(\boldsymbol{\omega}^{IL})$ has the smallest MSE only for one scenario (the same MSE as $L_1(1/N)$), its is always close to it.

Now we consider the scenarios with respect to their error variance similarity, i.e., basically how diversity of forecasting performances of the input forecasts. There are three variants, very similar ($z = 0.05$), less similar ($z = 0.20$) and dissimilar ($z = 0.50$) error variances of forecasts. Table 4.5 presents the percentages of how often each method has the best forecast accuracy and Table 4.6 depicts the average ranks and distances with respect to the error variances (and special groups).

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
$z = 0.05$	12.50	0.00	0.00	8.33	12.50	0.00	0.00	**83.33**	29.17	29.17	50.00
$z = 0.20$	0.00	0.00	0.00	8.33	0.00	0.00	8.33	33.33	45.83	**62.50**	25.00
$z = 0.50$	0.00	4.17	0.00	0.00	4.17	0.00	**58.33**	12.50	41.67	37.50	25.00

Table 4.5. Percentage of scenarios for which benchmarks, LHS and L_1 methods have the smallest MSE with respect to the error variance similarity (ex post analysis). The total number of scenarios for each error variance similarity is 24. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

Overall, the L_1 methods as a whole usually have the smallest MSE, smallest rank and distances for all error variance similarities.

The two L_1 methods that shrink towards a fixed value κ have opposing trends both with respect to the smallest MSE, average rank and distance. The more dissimilar forecasts are, the more often $L_1(0)$ has the smallest MSE (58.33% for $z = 0.50$). This can be due to the fact that in case of high error variance dissimilarity the method can more easily select a well suited set of forecasts. In contrast, for $L_1(1/N)$ the more similar forecasts are, the more often the method has the best forecast accuracy. Recall, the more similar forecasts are, the closer the true optimal solution is to equal weights, ceteris paribus.

Overall shrinkage towards prior weights, i.e., $L_1(\boldsymbol{\omega}^{E2})$ and $L_1(\boldsymbol{\omega}^{E3})$, oftentimes has the smallest MSE for $z = 0.20$ and $z = 0.50$, i.e., less similar error variances. For all z-values, the average rank of $L_1(\boldsymbol{\omega}^{E3})$ is slightly higher than for $L_1(\boldsymbol{\omega}^{E2})$ and $L_1(\boldsymbol{\omega}^{IL})$, except for $z = 0.05$. However, $L_1(\boldsymbol{\omega}^{IL})$ is the most consistent method over all error variance similarities again.

In addition to the average ranks and distance in Figure 4.13, the percentages of how often each method has the smallest MSE with respect to the special groups is shown in Table 4.6.

As a whole the L_1 methods are better than the benchmarks and LHS in terms of all percentage of being the best method, average rank and distance. However, Figure 4.13 shows that overall all L_1 methods, beside $L_1(0)$, have rather similar rank and distance

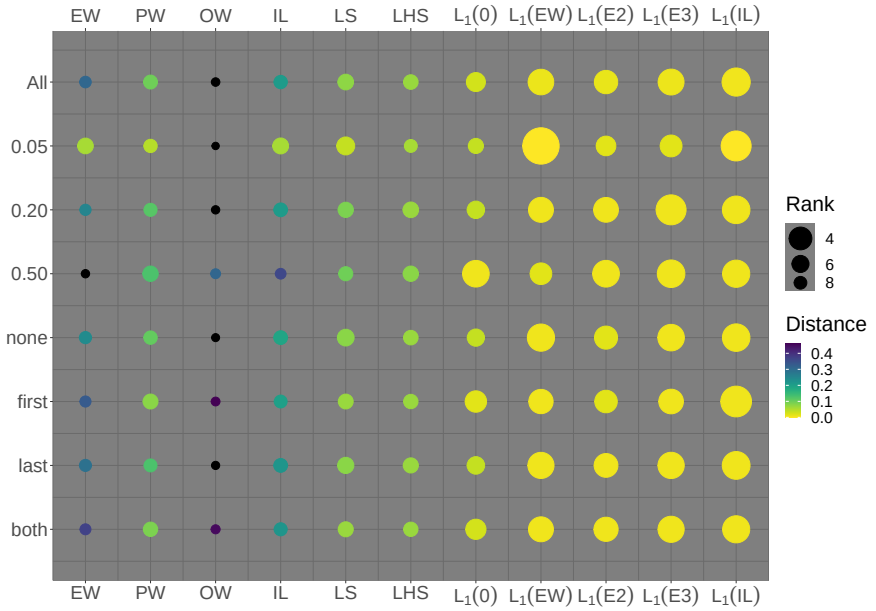


Figure 4.13. Illustration of average ranks and distances of the benchmark, LHS and L_1 methods for error variance similarities and special groups (ex post analysis).

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
none	11.11	0.00	0.00	5.56	11.11	0.00	11.11	**50.00**	38.89	33.33	27.78
first	0.00	5.56	0.00	5.56	0.00	0.00	33.33	33.33	33.33	38.89	**44.44**
last	5.56	0.00	0.00	11.11	11.11	0.00	11.11	50.00	44.44	**55.56**	44.44
both	0.00	0.00	0.00	0.00	0.00	0.00	33.33	38.89	38.89	**44.44**	16.67

Table 4.6. Percentage of scenarios for which benchmarks, LHS and L_1 methods have the smallest MSE with respect to special groups (ex post analysis). The total number of scenarios for each special group is 18. The methods with the highest percentages for each special group are depicted in bold numbers.

method-wise. The most noticeable exception is $L_1(\boldsymbol{\omega}^{IL})$ which has a smaller average rank and distance for SG first. Moreover, $L_1(0)$ and $L_1(1/N)$ have opposing trends, i.e., the former is better more often for SG first and both while the latter is better more often for SG none and last, see Table 4.6. For $L_1(0)$ this may be due the fact that it can more easily distinguish between which forecasts to chose if a particularly good group of forecast is present.

In general, the special groups have less of an effect on the forecasting performance than the structure of the error correlations (CM) and diversity of forecasts (z).

4.3.2 Out-Of-Sample: L_1 and LHS with Hyperparameter Estimation

In this section we consider the out-of-sample forecast accuracy of the LHS and L_1 methods with hyperparameter estimation. To this end, we use the exact same time series as in Section 4.3.1.[31] Recall, that for each time series we compute forecasts for the 50 observations in the test set based on a rolling window. However, previously in Section 4.3.1, we chose the best hyperparameter for each test set ex post. This is used similar in other research related to forecasting (see e.g., Radchenko et al., 2023). It provides value in the sense that we can analyze a potential forecast accuracy of the method if we would choose the best single parameter for the whole test set. By that, we remove the uncertainty hyperparameter estimation to focus on the potential performance. This approach differs in two ways from a real-world application.

First, the hyperparameter have to be determined a priori and not ex post. Second, oftentimes one estimates the hyperparameter for each observation of the test set individually. By that, the method can respond to current developments in the time series and adapt the model or method to produce the best possible forecast. Based on this, we are now considering a situation in which we have to forecast only based upon past or historic information but we re-estimate the hyperparameters for each observation in the test set.

To this end, we use cross-validation in its simplest form as discussed in Section 2.1. Recall, that we at its core it uses the ex post analysis approach of choosing the best hyperparameter, but only for the observations in the training set. To put it differently, we divide the training set into a training subset (75%, i.e., 30 observations) and a validation set (25%, i.e., 10 observations). Then we compute forecasts for candidate values of the hyperparameter for the validation set. Note that we use the same candidate value as in Section 4.3.1, i.e., from the smallest feasible value to 50 in 0.1 increments. The hyperparameter that minimizes the forecast accuracy in the validation set is used to forecast the next observation in the test set. This process is repeated for every observation in the test set. As a result we have 200 MSE values for each method and scenarios.

Tables 4.7 and 4.8 present the average MSE across the 200 test sets for each scenario. Both Tables 4.7 and 4.8 follow the same structure as Tables 4.1 and 4.2, which present the MSE result of the ex post analysis. The MSE values of the benchmark methods (EW, OW, PW and IL) in Tables 4.7 and 4.8 are identical to those from the ex post analysis follow Tables 4.1 and 4.2, because they do not have hyperparameters.

This is not the case for the LHS and L_1 methods. At first glance one can see that in contrast to the ex post analysis the MSE values of the LHS and L_1 methods for pseudo out-of-sample forecasting are not always superior to the benchmarks, see for example

[31] We want to emphasize that there is no usage of future information or insights based upon the ex post analysis to compute forecasts for the pseudo out-of-sample forecasting.

CM	z	SG	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\omega})$		
							LS	LHS	0	$1/N$	ω^{E2}	ω^{E3}	ω^{IL}
1	0.05	none	0.98	**0.97**	2.11	0.98	2.10	**0.97**	1.05	1.05	1.03	1.03	1.04
		first	0.97	**0.93**	1.89	0.96	1.88	0.94	0.99	1.01	0.99	1.00	1.01
		last	1.01	**0.99**	2.04	1.00	2.03	**0.99**	1.06	1.06	1.04	1.04	1.05
		both	0.97	0.93	1.75	0.96	1.74	**0.92**	0.97	0.98	0.96	0.96	0.98
1	0.20	none	1.16	0.92	1.03	1.12	1.03	0.83	0.70	0.72	**0.69**	0.70	0.71
		first	1.12	0.73	0.52	1.04	0.52	0.53	**0.36**	0.37	**0.36**	**0.36**	0.37
		last	1.24	0.95	0.84	1.18	0.84	0.76	0.57	0.58	**0.56**	**0.56**	0.57
		both	1.15	0.72	0.45	1.05	0.45	0.48	**0.31**	0.32	**0.31**	**0.31**	**0.31**
1	0.50	none	1.53	0.80	0.42	1.33	0.42	0.49	0.29	0.30	**0.28**	0.29	0.29
		first	1.38	0.30	0.09	0.85	0.09	0.11	**0.06**	0.07	**0.06**	**0.06**	**0.06**
		last	1.64	0.82	0.35	1.38	0.35	0.43	0.24	0.24	**0.23**	**0.23**	0.24
		both	1.46	0.30	0.08	0.85	0.08	0.10	**0.05**	0.06	0.06	0.06	0.06
2	0.05	none	**0.56**	0.62	1.37	**0.56**	1.37	0.59	0.67	0.62	0.63	0.62	0.62
		first	**0.56**	0.62	1.34	**0.56**	1.34	0.59	0.68	0.62	0.63	0.62	0.62
		last	**0.57**	0.64	1.37	**0.57**	1.37	0.60	0.69	0.62	0.64	0.63	0.62
		both	**0.57**	0.63	1.35	**0.57**	1.35	0.59	0.67	0.62	0.63	0.62	0.62
2	0.20	none	0.68	0.68	1.42	0.66	1.41	**0.65**	0.74	0.71	0.69	0.69	0.70
		first	0.64	**0.56**	1.11	0.60	1.10	**0.56**	0.60	0.64	0.60	0.59	0.62
		last	0.68	0.66	1.31	0.65	1.31	**0.62**	0.72	0.70	0.66	0.66	0.68
		both	0.66	**0.54**	1.00	0.60	1.00	0.55	0.58	0.62	0.57	0.57	0.60
2	0.50	none	0.87	**0.64**	1.10	0.76	1.10	0.66	0.65	0.71	**0.64**	0.65	0.68
		first	0.80	0.26	0.35	0.49	0.35	0.34	**0.23**	0.28	0.27	0.27	0.25
		last	0.96	0.66	1.05	0.81	1.04	0.67	0.66	0.70	**0.64**	**0.64**	0.67
		both	0.83	0.26	0.33	0.49	0.33	0.33	**0.22**	0.26	0.25	0.25	0.24
3	0.05	none	**0.25**	0.32	0.63	0.26	0.62	0.29	0.34	0.28	0.31	0.30	0.28
		first	**0.25**	0.31	0.63	**0.25**	0.62	0.28	0.33	0.27	0.30	0.29	0.27
		last	**0.25**	0.32	0.63	**0.25**	0.62	0.29	0.34	0.28	0.31	0.30	0.28
		both	**0.25**	0.32	0.64	**0.25**	0.63	0.28	0.34	0.27	0.30	0.29	0.27
3	0.20	none	**0.30**	0.36	0.71	**0.30**	0.70	0.32	0.39	0.33	0.35	0.34	0.32
		first	0.29	0.32	0.63	**0.27**	0.63	0.29	0.35	0.31	0.31	0.30	0.30
		last	0.30	0.36	0.71	**0.29**	0.70	0.31	0.38	0.33	0.33	0.32	0.32
		both	0.30	0.32	0.61	0.27	0.60	**0.28**	0.34	0.31	0.30	0.29	0.29
3	0.50	none	0.40	0.39	0.75	**0.35**	0.75	**0.35**	0.42	0.41	0.38	0.37	0.38
		first	0.35	**0.19**	0.35	0.22	0.35	0.24	0.21	0.26	0.24	0.24	0.22
		last	0.43	0.40	0.76	**0.36**	0.75	**0.36**	0.44	0.43	0.38	0.38	0.39
		both	0.37	**0.20**	0.36	0.22	0.36	0.24	0.21	0.27	0.25	0.24	0.22

Table 4.7. Simulation study results of benchmark, LHS and L_1 methods for correlation matrices CM1, CM2 and CM3 (out-of-sample analysis). The table depicts the MSE of the forecast combination method of Section 4.2 and other benchmarks. The methods with the smallest MSE are depicted in bold numbers.

$1/0.05/\cdot$. The estimation of the hyperparameters introduces estimation error into the methods. The fact that the LHS and L_1 methods have larger MSE values is even more prevalent because the MSE in the pseudo out-of-sample context can be smaller than the MSE of the ex post analysis. This is due to the fact that in the former we use different hyperparameters for all observations in the test set while for the latter we use a fixed hyperparameter value for all observations of the test set.

In what follows we will first analyze the results overall and then with respect to the correlation matrices, error variance similarities and special groups.

CM	z	SG	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
							LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
4	0.05	none	**0.63**	0.69	1.57	**0.63**	1.57	0.71	0.76	0.71	0.76	0.74	0.71
		first	**0.65**	0.70	1.61	**0.65**	1.61	0.71	0.76	0.71	0.75	0.73	0.71
		last	**0.66**	0.72	1.64	**0.66**	1.63	0.73	0.79	0.73	0.77	0.75	0.73
		both	0.64	0.68	1.59	**0.63**	1.59	0.70	0.74	0.70	0.73	0.72	0.69
4	0.20	none	0.79	0.78	1.78	**0.76**	1.78	0.79	0.86	0.84	0.85	0.83	0.83
		first	0.73	**0.67**	1.48	0.68	1.48	0.68	0.73	0.74	0.72	0.72	0.72
		last	0.80	0.78	1.74	**0.77**	1.74	0.78	0.87	0.85	0.85	0.84	0.84
		both	0.77	**0.68**	1.51	0.70	1.50	0.70	0.74	0.76	0.75	0.73	0.74
4	0.50	none	1.03	**0.79**	1.50	0.91	1.50	0.84	0.84	0.89	0.87	0.86	0.88
		first	0.90	0.30	0.45	0.56	0.45	0.43	**0.29**	0.31	0.31	0.31	0.31
		last	1.10	**0.79**	1.47	0.93	1.47	0.85	0.85	0.90	0.88	0.87	0.89
		both	1.02	0.31	0.46	0.60	0.45	0.44	**0.29**	0.33	0.32	0.32	0.32
5	0.05	none	0.44	**0.40**	0.84	0.45	0.84	0.50	0.44	0.43	0.47	0.46	0.44
		first	0.44	**0.40**	0.85	0.45	0.85	0.51	0.44	0.43	0.47	0.47	0.44
		last	0.44	**0.41**	0.88	0.46	0.87	0.52	0.44	0.43	0.48	0.47	0.44
		both	0.43	**0.41**	0.87	0.45	0.86	0.52	0.44	0.43	0.48	0.47	0.44
5	0.20	none	**0.51**	**0.51**	1.06	0.54	1.05	0.62	0.56	0.53	0.59	0.58	0.54
		first	**0.48**	0.49	0.98	0.50	0.97	0.57	0.53	0.50	0.56	0.55	0.51
		last	**0.52**	0.54	1.13	0.56	1.12	0.65	0.59	0.54	0.62	0.61	0.57
		both	**0.49**	0.53	1.06	0.51	1.05	0.58	0.57	0.52	0.58	0.57	0.54
5	0.50	none	0.67	**0.66**	1.21	0.69	1.20	0.74	0.72	0.70	0.75	0.74	0.71
		first	0.56	**0.31**	0.42	0.45	0.42	0.39	**0.31**	0.36	0.38	0.37	0.36
		last	**0.66**	0.67	1.22	0.68	1.22	0.72	0.73	0.71	0.74	0.74	0.71
		both	0.59	**0.33**	0.44	0.46	0.44	0.40	**0.33**	0.38	0.40	0.40	0.38
6	0.05	none	0.79	**0.63**	1.39	0.77	1.39	0.75	0.69	0.70	0.69	0.69	0.69
		first	0.78	**0.60**	1.32	0.75	1.32	0.72	0.65	0.67	0.66	0.66	0.66
		last	0.80	**0.63**	1.37	0.78	1.37	0.75	0.69	0.70	0.70	0.69	0.70
		both	0.80	**0.61**	1.30	0.77	1.30	0.71	0.66	0.68	0.67	0.67	0.68
6	0.20	none	0.97	**0.64**	1.16	0.89	1.16	0.72	0.65	0.68	**0.64**	0.65	0.67
		first	0.94	**0.52**	0.95	0.81	0.94	0.62	0.54	0.56	**0.52**	0.53	0.55
		last	1.02	0.64	0.96	0.91	0.96	0.67	0.62	0.62	**0.59**	**0.59**	0.62
		both	0.97	0.51	0.80	0.81	0.79	0.58	0.50	0.50	**0.47**	0.48	0.50
6	0.50	none	1.28	0.61	0.85	1.02	0.85	0.66	0.55	0.56	**0.52**	0.53	0.55
		first	1.22	0.26	0.42	0.62	0.42	0.41	**0.24**	0.25	**0.24**	**0.24**	0.25
		last	1.39	0.62	0.69	1.05	0.69	0.61	0.47	0.48	**0.44**	0.45	0.47
		both	1.31	0.26	0.38	0.62	0.38	0.39	0.23	0.24	**0.22**	0.23	0.23

Table 4.8. Simulation study results of benchmark, LHS and L_1 methods for correlation matrices CM4, CM4 and CM6 (out-of-sample analysis). The table depicts the MSE of the forecast combination method of Section 4.2 and other benchmarks. The methods with the smallest MSE are depicted in bold numbers.

4.3.2.1 Out-Of-Sample: Overall Results

Table 4.9 provides how often each method has the smallest MSE, the average rank and distance to the scenario-wise best method over all 72 scenarios. It is built in the same way as Table 4.3 in which we analyzed the ex post results. Furthermore, Figure 4.14 shows boxplots of the methods ranks over all scenarios. It is build in the same way as Figure 4.11. Figure 4.14 and the first row Table 4.9 show that the L_1 methods do not have a clear superiority over the benchmarks anymore.

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
Smallest MSE (%)	23.61	**36.11**	0.00	26.39	0.00	11.11	15.28	0.00	23.61	11.11	2.78
Avg Rank	6.49	**4.06**	9.92	5.42	9.54	5.39	5.35	5.48	5.12	4.62	4.61
Avg Distance	0.27	0.07	0.50	0.17	0.50	0.07	0.05	0.05	0.05	**0.04**	**0.04**

Table 4.9. Key figures for the MSE values of benchmark, LHS and L_1 methods over all simulation study scenarios (out-of-sample analysis). Smallest MSE (%) - Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank - Average rank of a method where a smaller rank is favorable. Avg Distance - Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

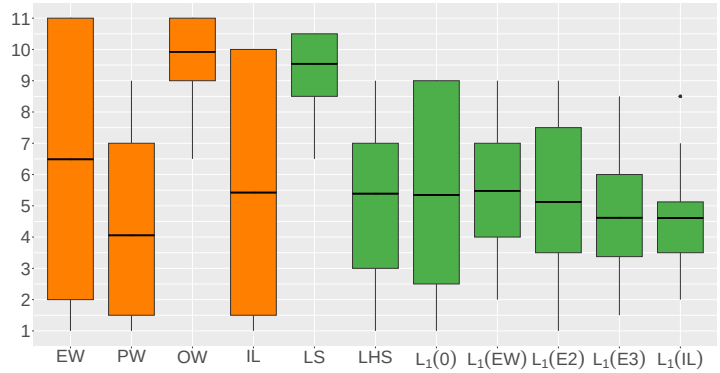


Figure 4.14. Boxplot of ranks across benchmarks, LHS and L_1 methods for the pseudo out-of-sample analysis.

While EW, PW and IL more often have the smallest MSE (Table 4.9) and, in case of PW, the smallest average rank. However, Figure 4.14 shows that the IQR, i.e., range of ranks, of both EW and IL is very large. Accordingly, for some scenarios they are competitive forecast combination methods while for others they are inferior to the other methods. Although PW also has a larger range or distribution of ranks than almost all L_1 methods, this also includes smaller ranks in comparison.

With respect to the LHS methods, interestingly, LHS is now favorable as one can clearly see by considering the generally larger ranks in Figure 4.14. This may be due to the fact that for LS we used the estimator provided by the authors Blanc and Setzer (2020) for the pseudo out-of-sample forecasting. For LHS there is no such estimator, i.e., we used the validation set approach. Note that for LS we also used the validation set approach for the ex post analysis. In the latter case we wanted to compare the methods based on the same future information. For the out-of-sample forecasting where only past information can be used, we decided to go with the way the authors intended.

For shrinkage towards a fixed value $L_1(\kappa)$, $L_1(0)$ has scenarios where it provided better ranks than other methods, however, this hold similar for other scenarios where it has noticeably higher ranks. The middle 50% of ranks for $L_1(1/N)$ is overall lower,

however, it does not reach the small average ranks $L_1(0)$ has. Nevertheless, it is more compact, i.e., it has a consistent forecast accuracy. For both $L_1(0)$ and $L_1(1/N)$ the average ranks (5.35 and 5.48) and distance to the best method (both 0.05) are similar.

The L_1 methods with shrinkage towards prior weights, $L_1(\dot{\boldsymbol{\omega}})$, have the smallest average ranks after PW. Additionally, all L_1 methods have smaller average distances, i.e., they are on average closer to the best method scenario-wise. $L_1(\boldsymbol{\omega}^{E2})$ also has the smallest MSE for about a quarter of scenarios. However, if we consider the middle 50% of ranks in Figure 4.14, we can see that they are larger compared to both $L_1(\boldsymbol{\omega}^{E3})$ and $L_1(\boldsymbol{\omega}^{IL})$. Basically, $L_1(\boldsymbol{\omega}^{E2})$ can provide better results for some scenarios, while $L_1(\boldsymbol{\omega}^{E3})$ and particularly $L_1(\boldsymbol{\omega}^{IL})$ provide competitive results more consistently.

In terms of consistency, shrinkage towards prior weights is superior to all other shrinkage methods. Evidence for this is both the average rank, distance, and the distribution of ranks depicted in Table 4.9 and Figure 4.14. Overall the positive weights approach both is the best method most often and has the smallest average rank. From the ex post analysis of Section 4.3.1 we know, that the L_1 methods can do better even if we use a fixed value. Therefore, we again analyze the forecast accuracy with respect to the underlying correlation matrix, error variance similarity and special groups to identify scenarios in which certain method are advisable to use.

4.3.2.2 Out-Of-Sample: Groups of Scenarios

In this section we analyze the forecast accuracy of the forecast combination methods with respect to the error correlation matrix, error variance similarity and special groups.

Table 4.10 and Figure 4.15 present how often each method has the smallest MSE, i.e., is the best method, and the average ranks and distances. The benchmark methods

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
CM1	0.00	25.00	0.00	0.00	0.00	25.00	33.33	0.00	**58.33**	41.67	16.67
CM2	**33.33**	25.00	0.00	**33.33**	0.00	25.00	16.67	0.00	16.67	8.33	0.00
CM3	41.67	16.67	0.00	**75.00**	0.00	16.67	0.00	0.00	0.00	0.00	0.00
CM4	25.00	33.33	0.00	**50.00**	0.00	0.00	16.67	0.00	0.00	0.00	0.00
CM5	41.67	**66.67**	0.00	0.00	0.00	0.00	16.67	0.00	0.00	0.00	0.00
CM6	0.00	50.00	0.00	0.00	0.00	0.00	8.33	0.00	**66.67**	16.67	0.00

Table 4.10. Percentage of scenarios for which benchmarks, LHS and L_1 methods have the smallest MSE with respect to the error correlation matrix (out-of-sample analysis). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

are superior more often for the correlation matrices that have less highly or mixed correlated forecast errors (CM 2, 3, 4, and 5). Shrinkage towards prior weights, $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$, we proposed to use with an L_1 constraint in Section 4.2.3 have the superior

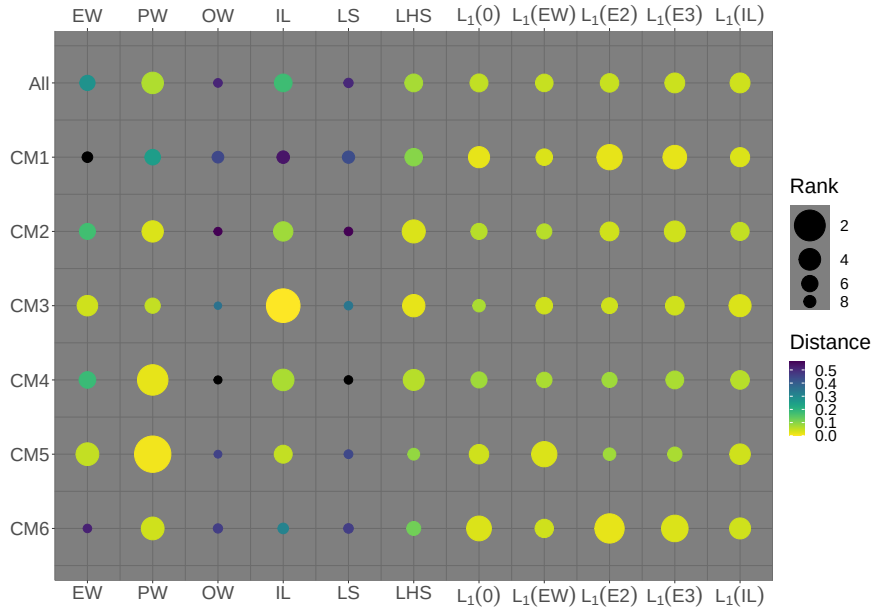


Figure 4.15. Illustration of average ranks and distances of the benchmark, LHS and L_1 methods for different correlations matrices (out-of-sample analysis).

forecast accuracy most often if highly correlated forecasts errors are common as in CM1 and CM6. Especially $L_1(\boldsymbol{\omega}^{E2})$, i.e., if the best 50% are shrunken and selected towards their equal weights while the rest is shrunken and selected towards zero with an L_1 constraint. Additionally, $L_1(\boldsymbol{\omega}^{E2})$ but also $L_1(\boldsymbol{\omega}^{E3})$ have the smallest average ranks for these correlation matrices (CM 1 and 6) as one can see in Figure 4.15.

Although $L_1(0)$ is the best method for some scenarios for almost all error correlation matrices, its average rank is always higher (smaller circle), i.e., for the scenarios where it is not, among others, the best method, it

With respect to CM2, although EW and IL have a smaller MSE more often, both PW and LHS have better average ranks (circle size) and distances (circle color), see Figure 4.15. For the correlation matrices CM3,4 and 5, the differences between the L_1 constraint methods are negligible in comparison to IL for CM3 (small error correlations 0.2) and i comparison to PW for CM4 (0.9 error correlation within and 0.5 between groups) and CM5 (better forecasts have higher error correlations). Although it is worth mentioning that for CM5 shrinkage towards a fixed value $L_1(\kappa)$ has better noticeably better ranks than shrinkage towards prior weights.

The following table depicts the percentage of how often a method has the smallest MSE for both the error variance similarities in the first three rows and special groups in the last four rows. Additionally, Figure 4.16 shows the average ranks and distances. The benchmarks methods are the best methods most often for $z = 0.05$, i.e., similar error variances. If one also takes Figure 4.16 into consideration, they are also very consistent (small average rank) and close to the best method (small average distance), the smaller

	EW	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
					LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
$z = 0.05$	**45.83**	**45.83**	0.00	**45.83**	0.00	12.50	0.00	0.00	0.00	0.00	0.00
$z = 0.20$	20.83	29.17	0.00	25.00	0.00	12.50	8.33	0.00	**33.33**	16.67	4.17
$z = 0.50$	4.17	33.33	0.00	8.33	0.00	8.33	**37.50**	0.00	**37.50**	16.67	4.17
none	27.78	**44.44**	0.00	27.78	0.00	16.67	0.00	0.00	27.78	0.00	0.00
first	22.22	**44.44**	0.00	22.22	0.00	5.56	33.33	0.00	22.22	16.67	5.56
last	27.78	22.22	0.00	**33.33**	0.00	16.67	0.00	0.00	27.78	22.22	0.00
both	16.67	**33.33**	0.00	22.22	0.00	5.56	27.78	0.00	16.67	5.56	5.56

Table 4.11. Percentage of scenarios for which benchmarks, LHS and L_1 methods have the smallest MSE with respect to the error variance similarity and special groups (out-of-sample analysis). The total number of scenarios for each error variance similarity is 24 and for special groups it is 18. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

the error variance similarity is. To a lesser extent this holds for LHS and $L_1(1/N)$. For all other L_1 methods the opposite is true. The more dissimilar the forecasts error variances are, the more often L_1 methods have the best MSE, smaller average ranks and distances. As a result, the shrinkage methods towards prior weights are particularly well suited for highly correlated forecast errors and more dissimilar, i.e., diverse forecasts.

This is most noticeable for shrinkage towards zero with $L_1(0)$. For very dissimilar forecasts it has the smallest rank, i.e., most consistent performance, over all methods by far. Taking look at special groups rows four to seven in Table 4.11 shows that the benchmark methods are most often the best methods for all special groups. This is most noticeable for the PW approach if we also consider Figure 4.16. The only shrinkage method that has a similar pattern is $L_1(\boldsymbol{\omega}^{E2})$ but to a lesser extent.

$L_1(0)$ method is useful if at least the first group is noticeably better, see both Table 4.11 and the average ranks in Figure 4.16. Again, this is probably due to the fact that if forecast are very dissimilar or groups are more distinguishable due to special groups , the method can select a favorable group of forecasts more easily. Going back to the results from Table 4.10 in conjunction with Tables 4.7 and 4.8 shows that the scenarios for which $L_1(0)$ has the smallest MSE is less dependent on the error correlation matrix but rather on the error variance similarity and special groups.

Other than that, no shrinkage method proves to be tailored towards a specific structure of special groups. However, overall the average ranks slightly decrease, i.e., methods are less consistent, if either no special group is present or if the last special group has a noticeably worse forecasting performance.

Overall the average ranks are similar to *All* at the top of Figure 4.16, i.e., if we do not differentiate between the special groups at all.

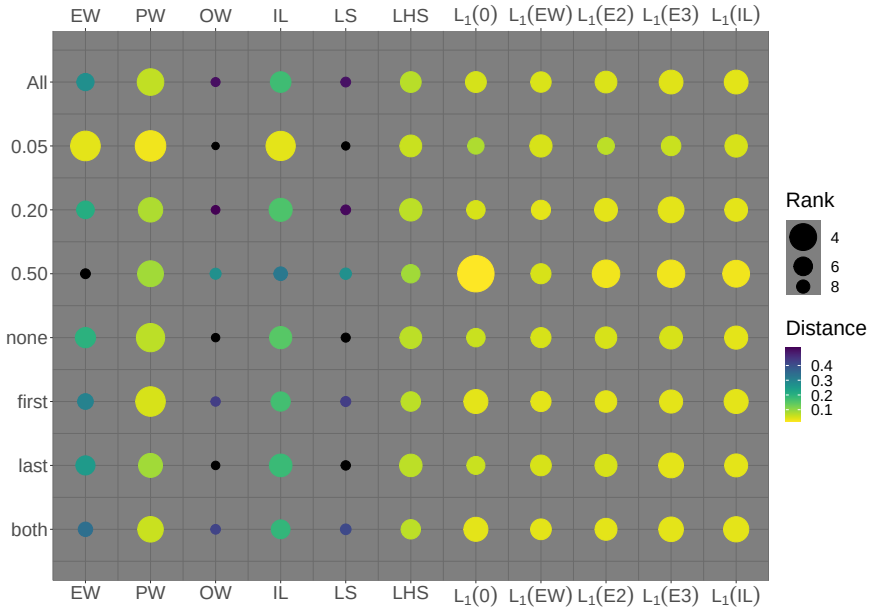


Figure 4.16. Illustration of average ranks and distances of the benchmark, LHS and L_1 methods for error variance similarities and special groups (out-of-sample analysis).

4.3.3 Summary of Results

In this section we briefly summarize the results of the simulation study.

For the ex post analysis, L_1 methods enhance the forecast accuracy compared to both the benchmark and LHS methods. Shrinkage towards equal weights as well as other, prior weights is favorable compared to shrinking all weights towards zero. $L_1(\boldsymbol{\omega}^{IL})$ is the most consistent method in the sense that it usually has small ranks over all scenarios. However, methods like $L_1(1/N)$ and $L_1(\boldsymbol{\omega}^{E3})$ can achieve smaller ranks for some scenarios but also larger for others, i.e., they have to be applied more situational than $L_1(\boldsymbol{\omega}^{IL})$. Overall, all L_1-methods beside $L_1(0)$ and to an extent $L_1(\boldsymbol{\omega}^{E2})$have similar small average ranks distance. Accordingly, they have consistently small MSE over all scenarios and if they are not the best method themselves they have a MSE close to it.

With regard to different correlations matrices, the L_1 methods as a whole have the smallest MSE more often, smaller average ranks and distance than the benchmark methods. For highly correlation correlated forecast errors $L_1(\boldsymbol{\omega}^{E2})$ and $L_1(\boldsymbol{\omega}^{E3})$ are superior. If inverse weights are used as a shrinkage direction, a consistently good forecast accuracy can be achieved over all correlation matrices. $L_1(1/N)$ is noticeably the best method if the error correlation increase with forecast accuracy (CM5).

For the different error variance similarities, again all L_1 constraint usually improve the forecast accuracy compared to the benchmark methods and LHS methods. If forecast are very similar, shrinkage towards equal weights is superior while for less similar and

dissimilar forecast shrinkage towards zero provides better results. Although, the more diverse forecasts are with respect to the forecast error variances, the more superior shrinkage towards prior weights particularly, $L_1(\boldsymbol{\omega}^{E3})$, is. It has higher percentage of being best and a better average rank. This holds similarly for $L_1(0)$, but the opposite is true for $L_1(1/N)$. It is superior if forecasts are less diverse, i.e., have similar error variances. This pattern is similar to that of EW to which $L_1(1/N)$ shrinks. This result is due to the fact that the more similar forecast are, the better EW is in general. If more dissimilar forecasts are present, i.e., also forecasts with higher error variances, those forecast are considered equally and, by that, the forecasting performance decreases.

The special groups do overall not affect the forecast combination methods. The only noticeably exception is $L_1(0)$. Its performance is better if a group of particularly good forecasts is present.

If we estimate hyperparameters within the out-of-sample analysis, the clear superiority of the L_1 (and LHS methods) over the benchmarks is not unambiguous anymore. With respect to the L_1 methods shrinkage towards prior weights either has the smallest MSE most often or is more consistent in its forecasting accuracy compared to other L_1 methods. For highly correlated forecast errors, shrinkage towards prior weights in form of $L_1(\boldsymbol{\omega}^{E3})$ is superior because it has the smallest MSE most often and the smallest average rank and distance.

For smaller error correlations (CM3) or mixed correlation matrices (CM4 and CM5), IL and PW are favorable, respectively. If forecast error variances are similar, the benchmarks methods (EW, PW and IL) are well suited for pseudo out-of-sample forecasting. The more dissimilar the error variances are, the better the L_1 methods except $L_1(1/N)$. For very dissimilar error variance the $L_1(0)$ method has the smallest average rank and distance.

With regard to special groups the result do not indicate that any methods is specifically tailored for any structure of special groups. The only exception is $L_1(0)$ which is better suited for scenarios where a group of particularly good forecasts is present.

4.4 Discussion and Future Work

In this chapter, we presented how shrinkage methods improve the forecast accuracy and how the L_1 constraint can be used for forecast combination. With respect to the overall structure of this thesis, this chapter includes the first three major contributions stated in Chapter 1: an extended framework for simulation studies, a unified framework for the L_1 constraint methods, and Conditional Group Equal Weights as a shrinkage direction. It demonstrates how forecast combination with constrained weights can improve forecast accuracy and, by that, contributes to answer the overarching research question: *how to further improve the forecast accuracy of a combined forecast using constrained weights?*

To this end, this chapter was centered around six objectives.

(I) In Section 4.1.1, we showed that the idea behind shrinkage is to alter the weight vector, i.e., shrink it, towards certain directions. We motivated this based on the LHS method that starts from the optimal weights solution and shrinks all or a subset of weights linearly towards equal weights and the complementary set of weights towards zero. In its simplest form it shrinks all weights from OW (no bias, larger forecast variance) to EW (bias, smaller forecast variance). Shrinkage is based around the idea of the bias-variance trade off depicted in Equation (4.1). The out-of-sample error variance can be decomposed into a squared bias term, a variance term of the forecast and an irreducible error. If the increase in (squared) bias introduced by constraint or shrinkage is more than offset by the reduction in variance, the resulting shrunken solution has a superior out-of-sample forecast accuracy. In other words, the solution that minimizes the empirical error variance is not identical to the solution which minimizes the actual error variance based on the true error variance covariance matrix, see again Figure 4.2.

(II) In Section 4.1.2 we showed how variants of the L_1 constraint are considered and implemented in various ways in the literature.

(II) / (III) After that in Section 4.2, we translated the discussed methods into a unified framework that minimizes a quadratic function, the error variance of the combined forecast, subject to linear constraints. Importantly, we incorporated the unity constraint for all methods. We showed how to transform optimization problems with L_1 constraints such that we can represent them with linear constraints. This is necessary to solve the optimization problems with the algorithm we use. In Section 4.2.1 we presented the optimization problem of Equation (4.28) which can be used to shrink weights towards any fixed value κ. This includes shrinkage towards zero ($L_1(0)$), i.e., weights are shrunken and selected towards zero, as used in Radchenko et al. (2023). Moreover, it can also represent the so-called egalitarian Lasso from Diebold and Shin (2019) which shrinks and selects weights towards equal weights ($L_1(1/N)$). In Section 4.2.2.2 we extended the optimization problem such that it allows shrinkage towards a prior weights vector ($L_1(\dot{\boldsymbol{\omega}})$), e.g., the inverse-loss average weight as used in Roccazzella et al. (2022). We showed that shrinkage towards prior weights nests shrinkage towards a fixed value as a special case.

Consequently, we presented a unified framework as a quadratic optimization problem with linear constraints in the sense of J. M. Bates and Granger (1969) by which one can implement and compare all considered L_1 methods (V).

(IV) In Section 4.2.3 we proposed to use shrinkage directions with the L_1 constraint we called *Conditional Group Equal Weights*. It is inspired by the LHS method (Schulz et al., 2022), the peLasso (Diebold & Shin, 2019), $L_1(0)$ and $L_1(1/N)$. It creates a sophisticated and more comprehensive hybrid between $L_1(0)$ and $L_1(1/N)$. For a predefined number of groups we assign a budget to each group that corresponds to the proportion of

the weight the group has for the smallest possible γ value, i.e., shrinkage parameter. Affiliation to a group is determined by a feature, for example the forecast accuracy. For the smallest possible value of the shrinkage parameter, each group members weight is the equal weight conditional to their budget. In other words, the conditional equal weight is the budget of a given group divided by the number of group members. First, we use two groups ($\boldsymbol{\omega}^{EW2}$), with a budget of one for group one (top 50% forecast accuracy) and zero for group two (bottom 50% forecast accuracy). This corresponds to LHS, i.e., we shrink a subset of weights from OW to their corresponding, conditional EW and the remaining weights from OW to zero. However, we use a L_1 constraint to shrink weights instead of linear shrinkage. By that, weights are not only shrunken but also selected to either the conditional equal weights or zero as the shrinkage intensity increases. Additionally, we propose to divide the forecasts in three groups based on their forecast accuracy. The first group has a budget of $2/3$, the second $1/3$ and the third $0/3$, i.e., we shrink and select weights towards two different conditional equal weights and zero. At its core our proposed shrinkage direction $L_1(\boldsymbol{\omega}^{EW2})$ and also to an extent LHS is a simplification of the partially egalitarian Lasso introduced by Diebold and Shin (2019). For a given shrinkage intensity the peLasso shrinks and selects weights to the conditional equal weights of non-zero weights within a solution and towards zero. Both $L_1(\boldsymbol{\omega}^{EW2})$,$L_1(\boldsymbol{\omega}^{EW3})$ and LHS simplify this idea by defining the number of weights to be shrunken towards zero and the conditional equal weights a priori.

As a byproduct of our analysis, we showed that peLasso can be implemented as a one-step procedure which was left for future research by Diebold and Shin (2019). To this end, we define a quadratic mixed integer optimization problem, see Appendix A. Because it is beyond the scope of this thesis, we leave it for future research to implement and evaluate this one-step procedure of the peLasso.

(VI) In Section 4.3 we conducted an extensive simulation study. It is based on the simulation study in Roccazzella et al. (2022). However, while they consider four scenarios we consider 72 different scenarios that are built around error correlation matrices, error variances and special groups. We analyzed the forecast combination methods both if hyperparameter are determined ex post (see e.g., Diebold & Shin, 2019; Radchenko et al., 2023) and if they are estimated repeatedly by cross-validation (see e.g., Radchenko et al., 2023; Roccazzella et al., 2022). The ex post analysis removes the uncertainty of hyperparameter estimation to assess a potential forecast accuracy. In the out-of-sample analysis we consider a more real-world procedure where hyperparameters are repeatedly estimate for each observation in the test set.

As a first finding, we observed that for the ex post analysis the LS method was slightly better than LHS. However, the opposite is true to a greater extent when it comes to out-of-sample forecasting. This may be due to the fact that for LS in the pseudo out-of-sample analysis we used the estimate of the shrinkage parameter provided by the

authors Blanc and Setzer (2020). In contrast, we used the validation set approach for the ex post analysis. For the ex post analysis, we wanted to compare the methods based on the same future information but for the out-of-sample forecasting where only past information can be used, we decided to go with the way the authors intended. However, there is no such estimate for LHS. Accordingly, future research should evaluate whether or not the LS method as a whole is less suited for out-of-sample forecasting compared to LHS or if it is due to the different ways of hyperparameter estimation.

The ex post analysis shows that L_1 methods are oftentimes superior compared to both benchmark and LHS methods in terms of forecast accuracy. Shrinkage towards prior weights, in particular the shrinkage directions we proposed to use $\boldsymbol{\omega}^{EW2}$ and $\boldsymbol{\omega}^{EW3}$, are advantageous if forecast error are highly correlated. Shrinkage towards zero ($L_1(0)$) is better suited for very dissimilar forecasts and if there is a group of particularly good forecasts present. In contrast, shrinkage towards equal weights ($L_1(1/N)$) has better result if there is less diversity in the forecasts error variances.

If hyperparameters are estimated rather than chosen ex post, the introduced uncertainty is large enough such that the clear superiority of the L_1 methods is no longer given. This is reinforced by the fact that due to the repeated re-estimation of the hyperparameter the smallest achievable MSE is smaller in the out-of-sample analysis compared to the ex post analysis where a fixed value is chosen for the whole test set. Without any further information, the PW method has the smallest MSE most often. L_1 methods that shrink towards prior weights ($L_1(\dot{\boldsymbol{\omega}})$) have the smallest MSE less often compared to PW. However, their average rank is closest to PW and, except for $\boldsymbol{\omega}^{EW2}$, they have a more consistent forecasting performance across different scenarios. For highly correlated forecast errors shrinkage towards prior weights $L_1(\dot{\boldsymbol{\omega}})$ with $\boldsymbol{\omega}^{EW2}$ and $\boldsymbol{\omega}^{EW3}$ (CGEW) is the best method most often and has a consistently good forecasting performance.

We suggest using the prior weights shrinkage directions (CGEW) if forecast error are highly correlated which, according to Winkler and Clemen (1992), is common in forecast combination problems. In contrast, for less correlated forecasts errors the benchmarks methods are a better choice, when it comes to out-of-sample forecasting. For very dissimilar forecast error variances, particularly in conjunction with a group of very good forecast, $L_1(0)$ provides promising results.

As an extension of our analysis, we encourage future research to compare all considered methods to double shrinkage via weighted least squares Lasso proposed by Liu et al. (2023). We briefly introduced the method in Section 4.1.2, however it was published so recently that we could not consider it in our analysis.

Due to the fact that the L_1 methods fall short of their potential forecast accuracy if hyperparameter are estimated, a closer examination of hyperparameter estimation is

needed. To this end, we one could use artificial neural networks that learn the structure of the error variances and covariances to estimate the best hyperparameter.

In this chapter, we presented various L_1 based methods within the same unified framework that can incorporates prior weights. Previously, the lasso-based methods used various L_1 constraints and implemented them differently with and without the unity constraint. In this chapter, we compared the discussed methods based on the same quadratic optimization problem with linear constraint, including a unity constraint, which has not been done previously. Moreover, we proposed to use a shrinkage direction, conditional group equal weights, for the L_1 constraint which turned out be one of the most favorable methods.

Lastly, we conducted a simulation study in which we analyzed the forecasting performance and how it changes for various designed scenarios, ceteris paribus. However, to asses the potential of any forecast combination methods a evaluation with real-world data is necessary. We will provide such an analysis in Chapter 7 where we compare all forecast combination methods considered in this thesis.

In conclusion, it is important to acknowledge the potential of simple benchmark methods. They can be highly effective to combine forecasts in certain scenarios, in particular in comparison to other methods that need to estimate hyperparameters. However, the incorporation of L_1 constraints that shrink weights towards prior weights is a valuable tool for improving forecast accuracy, and it is advisable to utilize this approach. To this end, it is essential to consider the structure of forecast errors, including error variances and error correlations, in order to identify the most appropriate methods for a given context.

5 Bounded Weights

In the previous chapter, we demonstrated how shrinkage or forecast combination with constrained weights can improve the out-of-sample forecast accuracy. To this end, we presented various methods that utilize an L_1 constraint to shrink and select towards a predefined shrinkage direction or prior weights. This includes shrinkage towards a fixed value as a special case. The L_1 constraints impose a restriction on the entire weight vector, $\boldsymbol{\omega}$, by introducing a budget. This budget defines the extent to which the weights can deviate from the prior weights. Consequently, some forecasts may have very large (small) positive (negative) weights, while others have weights closer to or exactly zero. Due to the high absolute weights, those input forecasts exert a greater influence on the combined forecast (in sense of a marginal effect). As a result, the combined forecast is more sensitive to changes in the input forecasts. Given the close relationship between forecast combination and regression models (Granger & Ramanathan, 1984), we will use the term marginal effect to emphasize the effect a forecast has on the combined forecast. If a forecast with a larger absolute marginal effect has a temporary inferior performance, it can affect the forecast accuracy of the combined forecasts to a greater extent. Recall, that part of the motivation for forecast combination is to diversify risk by using multiple forecasts. Although the L_1 constraint is generally beneficial in terms of forecast accuracy (see Chapter 4), it is not capable of mitigating large absolute weights for individual forecasts due to its constraints on the weight vector as a whole.

An alternative approach to reduce weights and diversify risk is to introduce bounds, i.e., feasible intervals, for weights. The positive weights (PW) approach provided in Equation (2.30) uses a lower bound of zero for each forecast's weight. Both lower and upper bounds are also used in the mathematically analogous problem of portfolio selection (Arratia, 2014, pp. 256-258). For forecast combination, Radchenko et al. (2023) extended the idea of the PW approach by imposing a lower bound other than zero to reduce the amount of negativity in the solution, while generally allowing it. They refer to this approach by "TR4" and describe it as a "one-step trimming approach" (Radchenko et al., 2023). Henceforth, we will refer to it as *LB* for lower bound. Note that by imposing a lower bound we implicitly also imposes an upper bound of $1-(N-1)$ times the lower bound, because the unity constraint has to be fulfilled. To illustrate, consider a scenario with $N = 4$ forecasts and a lower bound of -0.1. If three of four weights are equal to the lower bound, the largest weight that can be assigned to any forecast is 1.3. Otherwise, the solution is infeasible due to the unity constraint.

Consequently, this approach like the L_1 methods also allows for large absolute weights, i.e., forecasts with a larger absolute marginal effect.

We extend the approach of Radchenko et al. (2023) and propose to consider not only lower bounds, but also upper bounds. To illustrate, consider the previous example. If we impose an upper bound of 0.8 in addition to the lower bound of -0.1, at most two weights can be equal to the lower bound. The remaining two weights must sum to 1.2, which allows one weight to be equal to the upper bound and the other to be 0.4. Consequently, the weights have been shrunken and, by that, the marginal effect of each weight has been reduced. This leads to a more robust and diversified combined forecast.

Although the concept of limiting positive weights may appear less intuitive than constraining negative weights at first glance, it is precisely the upper bound that introduces new aspects to the problem. It nests existing forecast combination methods. One of these methods is the hardest benchmarks to beat when attempting forecast combination problems for decades resulting in the forecast combination puzzle discussed in Section 2.3: the equal weights forecast. Moreover, it also nests both the PW and OW approach as well as solutions in-between EW, PW and OW.

Let us briefly recapitulate and emphasize again the advantages and disadvantages of OW, PW and EW. The original forecast combination problem (OW), theoretically provides the error-variance minimizing, i.e., the best possible, weights (J. M. Bates & Granger, 1969). The combined forecast is unbiased if all input forecasts are unbiased. However, due the potential estimation error in the variance-covariance matrix combined with the sensitivity of weights with respect to those estimation errors, OW often has an inferior out-of-sample forecast accuracy. Constraining the solution space by limiting the amount of negativity either completely (PW) or partially by introducing a negative lower bound helps to improve the forecast accuracy. These lower bound constraints introduce bias into the solution. If the reduction in variance more than offsets the increase in (squared) bias, the out-of-sample forecasting performance of the combined forecast can improve, see again Equation (4.1). The equal weights (EW) forecast is characterized by a lack of estimation error as weights are predetermined rather than estimated. Accordingly, the equal weights forecast is biased. On the one hand, it diversifies risk by including all forecasts and weighting them equally and, thereby making it is more robust to outlier forecasts. On the other hand it is incapable of utilizing information from the training set to improve forecast accuracy (Aksu & Gunter, 1992; Blanc & Setzer, 2020; Chan & Pauwels, 2018; Graefe, Armstrong, Jones, & Cuzán, 2014; Radchenko et al., 2023; Smith & Wallis, 2009; Winkler & Clemen, 1992).

Our proposed approach of using both a lower and upper bound nests all three methods (EW, PW and OW) and solutions in-between them. By that, it can combine the advantages of the three methods while mitigating their flaws. Furthermore, the proposed bounded weights approach will be extended to incorporate prior information in form

of prior weights. This extension, *Forecast Combination with Bounded Prior Weights* (BW($\boldsymbol{\omega}$)), enables solutions between prior weights and OW.

The objectives for this chapter are the following:

(I) Introduce forecast combination with bounded weights

(II) Analyze the solutions, i.e., weights, that result from the introduction of lower and upper bound.

(III) Extend the approach of bounded weights by imposing bounds around prior weights, i.e., forecast combination with bounded prior weights

(IV) Analyze the performance of the methods within a larger scale simulation study for different scenarios both with and without hyperparameter estimation.

With respect to the overall structure of this thesis, the introduction of Forecast Combination with Bounded Weights is the fourth main contribution that we stated in Chapter 1. By that, we contribute to the overarching research question or objective of this thesis to improve the combined forecast by using additional constraints which, in this chapter, are bounds.

The remainder of this chapter is organized as follows. Section 5.1 introduces our new approach of Forecast Combination with Bounded Weights. Section 5.2 extends this idea by incorporating prior weights into the optimization problem. Lastly, Section 5.3 evaluates the forecasting performance of both the bounded weights and bounded prior weights approaches within our simulation study.

5.1 Between Identical and Individual Weights

In this section, first the bounded weights approach is introduced, and it is shown how the nested methods are part of the solution space. To this end, we consider the feasible values of the lower and upper bound in Section 5.1.1. Second, Section 5.1.2 analyzes the solutions in-between the nested method which have interesting properties. Lastly, Section 5.1.3 considers how to determine the lower and upper bound based on cross-validation. To this end, we derive an algorithm to reduce the computational burden that comes with the introduction of two hyperparameters (lower and upper bound).

5.1.1 Optimization Problem and Feasible Bounds

We extend the unconstrained optimization problem given in Equation (2.22) by adding both a universal or common lower and upper bound for the weights. To this end, in addition to the unity constraint we introduce N constraints that ensure that each weight of the N forecasts is smaller than a specified common upper bound $\overline{\omega}$. Similarly, we

define N additional constraints that ensure that each weight of the N forecasts is larger than a universal lower bound $\underline{\omega}$. The corresponding optimization problem is

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \omega_i \geq \underline{\omega} \qquad \forall\, i = 1, \dots, N, \\ & \omega_i \leq \overline{\omega} \qquad \forall\, i = 1, \dots, N \end{aligned} \tag{5.1}$$

By that, we define one common interval of feasible values for all weights, i.e., $\omega_i \in [\underline{\omega}, \overline{\omega}]$ $\forall\, i = 1, \dots, N$. The following figures depict two examples for the feasible interval based on different bounds.

(a) $\omega_i \in [-1.5, 1.5]$ **(b)** $\omega_i \in [-0.5, 1.5]$

Figure 5.1. Illustration of the lower and upper bound or interval of feasible values for the weights ω_i $\forall\, i = 1, \dots, N$.

Figure 5.1(a) showcases a symmetrical interval around zero for the feasible interval, i.e., the lower and upper bound have the same absolute value 1.5.[32] However, lower and upper bound do not have to have the same absolute value. Thus, we can have other intervals like for example the one depicted in Figure 5.1(b).

Based on the choice of lower bound $\underline{\omega}$ and upper bound $\overline{\omega}$ our approach nests the PW, OW and, most importantly, the EW approach. Within bounded weights we can introduce the EW in two different ways. First, as a consequence of using an upper bound. Second, EW is also introduced into the solution space if one uses a different range of feasible values for the lower bound than Radchenko et al. (2023).

Feasible Values of Lower and Upper Bound Table 5.1 presents the ranges of feasible values for both the lower bound (rows) and upper bound (columns) and illustrates for which combination of the lower and upper bound the solution of the BW approach is identical to other commonly used forecast combination approaches.

Due to the unity constraint, the smallest feasible upper bound is $\overline{\omega} = 1/N$. Any smaller upper bound results in an empty solution space as the unity constraint in Equation (5.1) is violated. Accordingly, for the upper bound it has to hold that

$$\overline{\omega} \in [1/N, \infty). \tag{5.2}$$

This range is depicted by the columns in Table 5.1.

[32] For a formal description of the figure recall Figure 4.5.

$\underline{\omega}$ \ $\overline{\omega}$	$1/N$	...	∞
$1/N$	EW	...	EW
⋮	⋮	⋱	⋮
0	EW	...	PW
⋮	⋮	⋱	⋮
-∞	EW	...	OW

Table 5.1. Feasible values for the lower and upper bound and nested methods.

For the feasible value of the lower bound Radchenko et al. (2023) use $\underline{\omega} \in (-\infty, 0]$. By that, they restrict negative weights to not become too small.[33] However, they dismiss positive values of the lower bound which are not only feasible but also sensible. A strictly positive lower bound forces a solution in which every forecast is considered with a positive weight, like for example in the equal weight solution. Again due to the unity constraint, the largest feasible value of the lower bound $\underline{\omega}$ is $1/N$, i.e.,

$$\underline{\omega} \in (-\infty, 1/N]. \tag{5.3}$$

The Solution Space To analyze the solution space of our method, we further consider Table 5.1.

First, if we use $1/N$ either for the lower or upper bound or both, the solution will always be the equal weights solution regardless of the other bound. Any deviation of weights from this solution towards either negative infinity or positive infinity requires at least one weight to increase or decrease respectively to offset the deviating weights and fulfill the unity constraint. However, this offset is impossible due to the proposed lower or upper bound being $1/N$. The first column of Table 5.1 shows this for the upper bound, and the first row for the lower bound respectively.

Second, we focus only on the lower bound as proposed by Radchenko et al. (2023), i.e., only non-positive values for the lower bound, and assume that there is no upper bound, i.e., $\overline{\omega} \to \infty$ or the corresponding constraint is simply omitted from Equation (5.1). The approach by Radchenko et al. (2023) is depicted in the last column for $\underline{\omega} \in (-\infty, 0]$ in the orange-shaded area. If the lower bound is set to zero, the proposed optimization problem is identical to the PW approach. As the lower bound decreases towards negative infinity, i.e., $\underline{\omega} \to -\infty$, for a sufficiently small value of $\underline{\omega}$, we denote as $\underline{\omega}^*$, the lower bound no longer constrains the solution space. At this point, the optimization problem becomes the original forecast combination problem, OW, of Equation (2.22).[34] If only

[33] Recall that in the context of negative weights, smaller weights correspond to larger absolute weights, i.e., weights with a larger absolute marginal effect.

[34] Recall that we already used this concept in Chapter 4 with γ^*.

a lower bound is used with values between $\underline{\omega} \in (-\infty, 0]$, it bridges the gap between the PW and OW approach. It creates a *transition* between or *path* $\overleftrightarrow{PO}$ that connects the PW and OW approach. Henceforth, we define those *transition* or *paths* by the first letters of the corresponding approaches. Solutions in-between PW and OW can have aspects or properties of both approaches. However, depending on the value of the lower bound, the solution can gravitate more towards either the PW or OW solution. A lower bound closer to zero forces a solution with less negative weights, i.e., it is closer to the PW approach. In contrast, a smaller lower bound puts fewer limitations on negative weights which leads to a solution closer to OW.

Third, consider again the lower bound but where $\underline{\omega}$ is positive, i.e., $\underline{\omega} \in [0, 1/N]$. Note that there is still no upper bound. This case is depicted in the last column of Table 5.1 by the blue-shaded area and the cell of Table 5.1 that holds PW. Recall that if $\underline{\omega} = 1/N$ the only feasible solution is the equal weights and if $\underline{\omega} = 0$ we have the PW solution. Accordingly, as the lower bound is decreasing from $1/N$ towards zero, the solution shifts away from the EW solution towards the PW solution. By that, we create a path from EW to PW, i.e., $\overleftrightarrow{EP}_{lb}$. Note that the subscript is used to distinguish between two transitions that exist from EW to PW. In the presented case the path $\overleftrightarrow{EP}_{lb}$ is caused by extending the interval of feasible value from Radchenko et al. (2023) and allowing for positive weights for the lower bound (lb). The whole path from EW over PW to OW will be referred to as $\overleftrightarrow{EPO}$. It includes both $\overleftrightarrow{EP}_{lb}$ and $\overleftrightarrow{PO}$.

Fourth, let us consider our proposition to extend the optimization problem through the incorporation of the upper bound. By that, the EW solution is again introduced into the solution space. However, more importantly, the upper bound creates an additional, different path between the EW and PW solution ($\overleftrightarrow{EP}_{ub}$) and also a new transition from EW to OW ($\overleftrightarrow{EO}$). Based on the value of the upper bound, the in-between solution can either gravitate more towards EW or the PW or OW solutions respectively. Both transitions $\overleftrightarrow{EP}_{ub}$ and $\overleftrightarrow{EO}$ are depicted in Table 5.1 within the green-shaded area. The transition between the EW and PW is given for $\underline{\omega} = 0$. With an increasing upper bound the solution moves away from the equal weights solution and converges towards the positive weights solution. How this path, $\overleftrightarrow{EP}_{ub}$, is different from the one between PW and EW, $\overleftrightarrow{EP}_{lb}$, will be discussed in Section 5.1.2. The path between EW and OW, $\overleftrightarrow{EO}$, is a result of the lower bound $\underline{\omega} \to -\infty$, i.e., if the lower bound is omitted or sufficiently small and only an upper bound is used. For an increasing upper bound the solution transitions from the equal weights to the optimal weights solution.

Lastly, our proposed approach also allows transitions or paths between intermediate solutions. For that we chose a value of the lower or upper bound from the defined ranges in Equations (5.2) and (5.3) and vary the corresponding other bound. For example, we can choose a lower bound between zero and negative infinity such that the lower bound is still constraining the solution space, i.e., we do not allow for the transition $\overleftrightarrow{EO}$ described

earlier. By that, we create a transition starting at EW on the left-hand-side of Table 5.1 and then transitioning towards an intermediate solution somewhere between PW and OW as the upper bound is increased. Likewise, let us consider a value for the upper bound larger than $\frac{1}{N}$ but sufficiently small such that the upper bound still constrains the solution space. As the lower bound is decreased from $\underline{\omega} = 1/N$ towards $-\infty$, the solution starts in the first row of Table 5.1 with the EW solution as $\underline{\omega} = 1/N$. Then it transitions through a solution between equal and positive weights as $\underline{\omega} = 0$ and lastly moves towards an intermediate solution of EW and OW when $\underline{\omega} \to -\infty$.

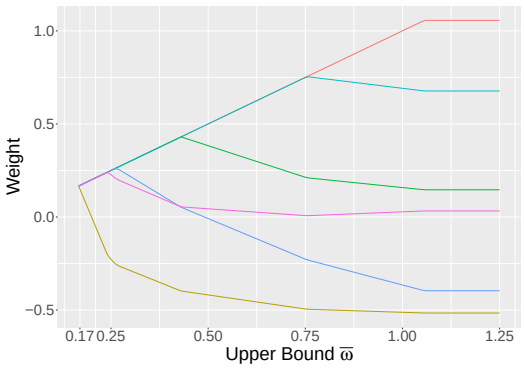
In this section we presented the forecast combination problem with bounded weights as well as the feasible values of the lower and upper bound and showed how bounded weights nests EW, PW and OW. In what follows we will visualize and further discuss the solutions of the bounded weights approach along the paths between the nested methods.

5.1.2 Solutions with Bounded Weights

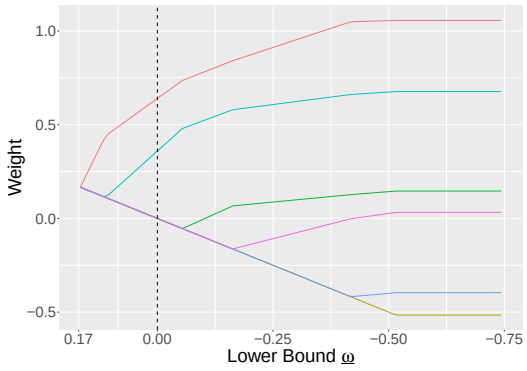
In order to obtain a better understanding of solutions using bounded weights, we analyze the transitions discussed in the previous Section 5.1.1 in more detail based on simulated data. This concerns both the transitions between the well-known EW, PW and OW approaches as well as between intermediate solutions. To this end, we use the same data set as in Chapter 4. Recall that it was generated using the simulation study proposed in Section 3.2 for the first correlation matrix $CM1$, i.e., highly correlated forecast errors, a relative group distance of $z = 0.5$ and no special group. Although, we simulated $N = 24$ forecasts, for the sake of simplicity six forecasts where chosen randomly. Then weights are calculated for feasible values of the lower and upper bound as defined in Equations (5.2) and (5.3). Figure 5.2 depicts the weights of individual forecasts (ordinate) depending on the lower or upper bound (abscissa), respectively. Each color represents the weight of a forecast. Recall that throughout this thesis, including both Figures 5.2 and 5.3, we use the same color for the same forecast.

Between the nested Methods Let us first consider the transition $\overleftrightarrow{EO}$ which is the result of having an upper bound $\overline{\omega}$ but no lower bound, i.e., $\underline{\omega} \to \infty$. With respect to Table 5.1 the figure presents the last row, i.e., from the equal weights solution to the optimal weights solution. Figure 5.2(a) shows the weight on the ordinate for each of the six forecasts for different values of the upper bound $\overline{\omega}$ depicted on the abscissa.

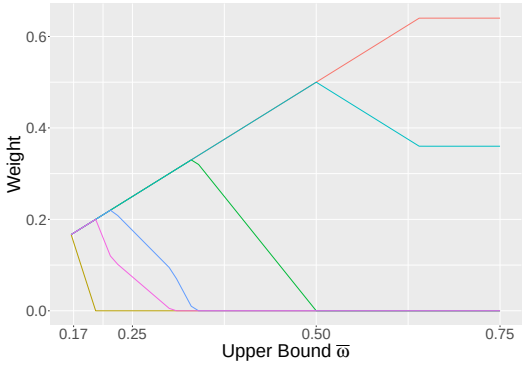
Starting on the left side of Figure 5.2(a) at the smallest feasible lower bound $\underline{\omega} = 1/N$, the optimal and only solution are equal weights. With increasing $\overline{\omega}$ the majority of weights is increasing by the same amount, i.e., they have *identical weights*. In turn this increase must be offset by at least one decreasing weight (yellow). To put it differently, at least one forecast is assigned a different *individual weight*. As $\overline{\omega}$ increases further



(a) Equal weights to optimal weights (no lower bound), i.e., $\overleftrightarrow{EO}$.



(b) Equal weights over positive weights to optimal weights (no upper bound), i.e., $\overleftrightarrow{EPO}$.



(c) Equal weights to positive weights ($\underline{\omega} = 0$), i.e., $\overleftrightarrow{EP}$.

Figure 5.2. Illustration of forecast weight paths for Bounded Weights between benchmark methods. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

this pattern continues. However, the number of forecasts having identical weights is gradually decreasing, i.e., more forecasts get individual weights, until all forecasts have individual weights. In Figure 5.2(a) after the yellow weight, first the pink, then the blue et cetera leave the identical weight group and get an individual weight. At about $\overline{\omega} = 0.75$ all weights have individual weights. For roughly $\overline{\omega} \geq 1.1$ the solution is identical to the OW solution.

This is of course only one example of the potential transition $\overleftrightarrow{EO}$, but it showcases an appearance of it. There are, however, other paths. For example, weights start positive, then become negative for certain values of the upper bound and, then, again have a positive value as the upper bound increases further. Moreover, it is possible that two weights have the same, unconstrained weight in the OW solution and, thus, technically not all weights are *individual*. However, the *identical* weights of the transition $\overleftrightarrow{EO}$ are caused by the upper bound which can be verified by considering the OW solution.

For the sake of transparency we included eight more such transitions in Figure B.1 which is located in Appendix B. It shows the transition $\overleftrightarrow{EO}$ for different scenarios of our simulation study as well as two real world time series taken from a well-known collection of time series, the M4 data set, which will be introduced in Chapter 7.

To better understand how our method works and why weights are chosen in this way, it is important to keep in mind what happens when we are imposing an upper (or lower) bound to our optimization problem: the solution space is constraint. Weights are forced to deviate from the optimal weights, OW, solution. Accordingly, as the upper bound is increased weights deviate more from equal weights strive towards their OW values as much as they are allowed to. As a result there is a group of forecasts with identical weights. For a larger upper bound, weights leave this identical group and get individual weights. The individual weights leave the identical group at some point and then strive towards the OW, because it minimizes the in-sample error variance of the combined forecast under the given constraints.

Lastly, within the solution of $\overleftrightarrow{EO}$, weights are usually non-zero, i.e., they contribute to the combined forecasts and, by that, we diversify risk. This is further supported by the fact that for smaller upper bounds the weights have similar marginal effect, i.e., they contribute similarly to the combined forecast. Additionally, in accordance to the linear shrinkage (LS) method discussed in Chapter 4 we have a solution between equal weights (high bias, low variance) and optimal weights (unbiased, high variance). By imposing constraint we trade bias for variance which can improve the out-of-sample forecast accuracy, recall the bias-variance trade-off of Equation (4.1).

If only a lower bound is used the solution paths have overall a certain similarity to the case where only an upper bound is used. Figure 5.2(b) shows the path $\overleftrightarrow{EPO}$ represented in last column of Table 5.1. It starts at the EW, transitions towards the PW (black dashed line at $\underline{\omega} = 0$) and ends at the OW solution. The abscissa in Figure 5.2(b) presents values of the lower bound $\underline{\omega}$. Note that the abscissa is reversed, i.e., it starts on the left-hand side at $1/N$ and then decreases. Similarly to the $\overleftrightarrow{EO}$ transition of Figure 5.2(a) the solution starts at equal weights on the left-hand side. However, as the lower bound decreases, one individual weight (red) increases while all others decrease by the same amount again resulting in an identically weighted set of forecasts. For a smaller value of the lower bound another weight (blue) deviates from this identical weighted set and gets an individual weight. This further proceeds until all forecasts have an individual weight and the solution reaches the OW solution for roughly $\underline{\omega} \geq 0.52$. Similar to the $\overleftrightarrow{EO}$ path two sets or groups of forecast weights are present within $\overleftrightarrow{EPO}$: the identical and individual weights. However, the identical weights are now *decreasing identical weights* compared to the *increasing identical weight* from $\overleftrightarrow{EO}$. In summary, while the upper bound diversifies risk by constraining weights with larger absolute weights, i.e., larger absolute marginal effect, the lower bound does the same by

constraining weights with very small weights, i.e., larger absolute marginal effect, upon the combined forecast.

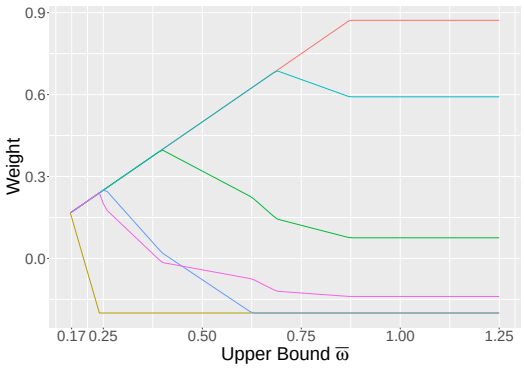
Considering the right-hand side of both Figures 5.2(a) and 5.2(b) one can see that, by definition, they transition to the same OW solution. However, the weights for the same forecast take different routes to this OW solution. This is well evident by the weight of the forecast depicted in green. Where in Figure 5.2(a) the weight starts and stays positive and then transitions to the OW solution from above, in Figure 5.2(b) the weight for the same forecast while starting positive, becomes negative shortly and then strives towards the OW solution from below.

In summary, we see a similar but mirrored behavior of the $\overleftrightarrow{EO}$ and $\overleftrightarrow{EPO}$ paths in terms of forecasts weighted identically while others get individual weights. Again, for the sake of transparency, we include eight more transition $\overleftrightarrow{EPO}$ in Appendix B Figure B.3.

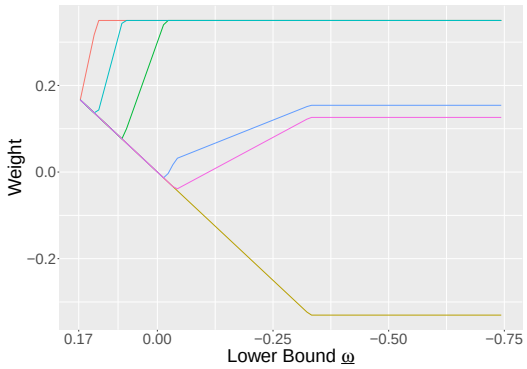
Lastly, Figure 5.2(c) shows the path of weights (ordinate) for different values of the upper bound (abscissa) given that the lower bound is $\underline{\omega} = 0$, i.e., the path between equal and positive weights ($\overleftrightarrow{EP}_{ub}$). The upper half of the figure looks similar but not identical to the weight paths shown in Figure 5.2(a). Starting from equal weights, weights again increase by the same amount, while some (yellow, purple, blue et cetera) get individual weights that are not constraint. In contrast, for $\overleftrightarrow{EP}_{ub}$ most of the individual weights first get smaller and eventually become zero instead of negative. The non-zero weights (teal and red) transition towards the PW solution.[35] Again, eight more examples for ($\overleftrightarrow{EP}_{ub}$) are given in Appendix B Figure B.2.

Beside $\overleftrightarrow{EP}_{ub}$ using the lower bound there is also a path between EW and PW which is part of $\overleftrightarrow{EPO}$ depicted in Figure 5.2(b), however, it is different. We referred to the latter one by $\overleftrightarrow{EP}_{lb}$. The transition $\overleftrightarrow{EP}_{lb}$ is part of Figure 5.2(b), starting from the left-hand side at equal weights up to the vertical, dashed black line at $\underline{\omega} = 0$. At this point, both $\overleftrightarrow{EP}_{ub}$ (Figure 5.2(c) with $\overline{\omega} \to \infty$) and $\overleftrightarrow{EP}_{lb}$ (Figure 5.2(b) with $\underline{\omega} = 0$) have the same solution, two non-zero weights (red and teal) while all others are zero. However, the $\overleftrightarrow{EP}_{ub}$ depicted in Figure 5.2(c) transitions between the equal weights solution and the positive weights solution by having a group of increasing identical weights from which individual weight deviate and become zero. In contrast, the $\overleftrightarrow{EP}_{lb}$ depicted in Figure 5.2(b) has a group of decreasing identical weight which, collectively, becomes zero at $\underline{\omega} = 0$, while the two forecasts that are part of the PW solution have individual weights.

[35] Note that this is the only set up where one can argue that a variable selection is performed. This is the case if less than N weights are non-zero in the PW solution, since for $\overline{\omega} = 1/N$ all N weights are non-zero.



(a) Weight path for varying upper bound given that $\underline{\omega} = -0.2$.



(b) Weight path for varying upper bound given that $\overline{\omega} = 0.35$.

Figure 5.3. Illustration of forecast weight paths for Bounded Weights for $\underline{\omega} = -0.2$ and $\overline{\omega} = 0.35$. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

Intermediate Solutions In the previous paragraph bounded weights were discussed with special attention to the nested benchmark methods to gain insights into how the bounds affect the weights. We either omitted one bound or, in case of $\overleftrightarrow{EP}_{ub}$ fixed the value of the lower bound at $\underline{\omega} = 0$. However, one can of course fix one of the bounds at any other value and conduct a similar analysis of the resulting weight path. Basically, comparing Figures 5.2(a) and 5.2(c) provides a good indication on how this may look like. Nevertheless, for the sake of completeness we depict such weight paths in Figure 5.3. In Figure 5.3(a) the abscissa represents the values of the upper bound. The lower bound is fixed at $\underline{\omega} = -0.2$. For an increasing upper bound, the paths exhibit a behavior similar to $\overleftrightarrow{EP}$ (Figure 5.2(c)). In particular, there is a group with increasing identical weights while gradually some weights (yellow, purple, blue et cetera) deviate from this group as the upper bound increases, i.e., they get individual weights. At some points some weights are equal to the lower bound (blue and yellow) which was also true for $\overleftrightarrow{EP}$ with a lower bound of zero, however there are more weights at the lower bound (yellow, purple, blue, and green). The solutions depicted in both Figures 5.2(c) and 5.3(a) can consist of three groups. The first two are the identical weights and the last group are the individual weights. For example, looking at the solution in Figure 5.3(a) for $\overline{\omega} \approx 0.6$, there is the identical weight set (teal and red), the individual group (green and pink) and a third group which also has identical weights (yellow and blue). Basically, we have a set of identical weights caused by the upper bound (*ub identical*) and a set of identically weighted forecasts caused by the lower bound (*lb identical*). The latter group is also present within the $\overleftrightarrow{EP}$ path depicted in Figure 5.2(c) with a constant equality weight of zero.

Lastly, let's focus on the weight path in Figure 5.3(b) which shows the lower bound on the abscissa, again decreasing from left to right. The upper bound is set to $\overline{\omega} = 0.35$. To an extent it is similar to Figure 5.2(b) which shows the path $\overleftrightarrow{EPO}$. Starting from equal weights, there is a lb identical weights group, that becomes negative for $\underline{\omega} < 0$. Other weights deviate from this set prior to this point and get an individual weights, e.g., red at $\underline{\omega} > 1/6$, teal at about $\underline{\omega} = 0.15$ or the green weight at roughly $\underline{\omega} = 0.08$. Those three weights deviate from the set of lb identical weights, to become individual temporarily before being constrained by the defined upper bound. The remaining weights (blue, purple, and yellow) all get an individual weight at some point.

In summary, based upon the insights from both Figures 5.2 and 5.3, for a given pair of lower and upper bound, the solution of our proposed bounded weights approach can consist of up to three groups of forecasts or weights. Two sets of identically weighted forecasts and a set of individually weighted forecasts. We can differentiate between them by the constraining bound, i.e., *ub identical* if caused by the upper bound and *lb identical* if the lower bound is responsible. If there is one or two sets of identically weighted forecasts we can trivially recognize by which bound they are caused. The bounds are known for a given solution and the *ub identical* weights are equal to $\overline{\omega}$ while weight from the *lb identical* set are equal to the lower bound. Note that there is the case where there can be three sets of identically weighted forecasts, if two weights that we characterize as individual cross, as it is the case for the purple and blue line in Figure 5.3(a). However, *ub identical* is simply the set with the largest identical weights while *lb identical* has the smallest identical weights.

Possible Group Constellations In the previous examples in Figures 5.2 and 5.3 all possible weight allocations of the BW approach were already present. However, those figures and the analysis focused on the transitions or paths of weights. An actual solution for specific values of the lower and upper bounds are given at the abscissa. It can be visualized as a vertical (black dashed) line like in Figure 5.2(b). To conclude this section, we briefly summarize the possible combinations of identical and individual weight sets. To this end, Table 5.2, depicts all possible six scenarios or combinations that can occur for any given time series. A check mark indicates that the set, shown by the columns, is part of the solution. The following list describes each row in the same order as Table 5.2 and refers to the figure where this constellation is present within our examples:

- The solution includes both an ub and lb identical set as well as individual weights. An example is given in Figure 5.3(a) for $\overline{\omega} = 0.625$.
- There is a lb identically and individually weighted set of forecasts, e.g., Figure 5.2(b) for $\underline{\omega} = -0.25$.

ub Identical	Individual	lb Identical
✓	✓	✓
–	✓	✓
✓	✓	–
–	✓	–
✓	–	–
–	–	✓
✓	–	✓

Table 5.2. Possible weight allocations or combinations of the two identical and one individual set for Bounded Weights.

- Similarly, there is an ub identically and individually weighted set of forecasts, e.g., Figure 5.2(a) for $\overline{\omega} = 0.5$.
- There are only individual weights, i.e., no two weights are either equal to the lower or upper bound, e.g., Figure 5.2(a) for $\overline{\omega} = 1$.
- Rows five and six depict the case where the BW solution is equal weights, i.e., either $\overline{\omega} = 1/N$ or $\underline{\omega} = 1/N$. Examples are trivially Figure 5.2(a) and Figure 5.2(b) respectively.
- The last row shows the last, theoretically possible constellation of two groups of identically weighted forecast. This constellation has not been observed within the considered examples.

5.1.3 Hyperparameter Determination

Our proposed approach of forecast combination with bounded weights involves two hyperparameters. Therefore, we need to determine both a value for the lower bound $\underline{\omega}$ and upper bound $\overline{\omega}$, if it is used to forecast unknown future values. The hyperparameters could be determined based on prior beliefs and information or in accordance to one's preference. Let us consider two examples for this. First, negative weights are allowed but limited using $\underline{\omega} = -0.1$, while positive weight are less constrained with $\overline{\omega} = 1$. Second, we want to ensure that only positive weight are used, but they should not be too large. Furthermore, all forecasts participate in the combined forecast to a minimum degree, i.e., $\underline{\omega} = 0.1$ and $\overline{\omega} = 0.5$. Alternatively, one can use a data driven approach, i.e., cross-validation. Recall that, we use a proportion of the training set as the validation set. We then forecast the observation of this validation set for a defined search grid of lower and upper bounds. Then we choose the hyperparameters that have the smallest MSE in the validation set.

To this end, we need to define the search grid, i.e., candidate values for the lower and upper bound. Recall that for the L1 constraint methods in Section 4.2 there is only one hyperparameter γ. For the BW approach we need to define two search grids, one for each bound. As a result, we get a matrix with candidate pairs of lower and upper bounds that we have to consider. This matrix is illustrated by Table 5.1. Assume that we use the same number of candidate values P both for the search grid of the L_1 approaches and for the two search grids (lower and upper bound) of the BW approach. Then the number of candidate pairs of BW is P^2 compared to P candidate values for L_1 methods. However, at least for our current implementation in R (R Core Team, 2022), solving BW for a single candidate pair is faster than solving L_1 methods. Nevertheless, the number of candidate pairs for BW is, ceteris paribus, larger. Thus, it is important to reduce the number of candidate lower and upper bounds as far as possible without omitting meaningful candidate pairs.

In order to define the search grids for the lower and upper bound we need to define the smallest and largest value. For methods with an L_1 constraint we can determine the smallest and largest γ values by Equations (4.37) and (4.38). For L_1 methods we can determine the largest possible value γ^* for a given data set. However, it is not straightforward for bounded weights as the largest/smallest possible value of one bound depends on the current value of the other bound. In what follows, we will show how to determine the end point of a search grid for a given data set in Section 5.1.3.1. Thereafter, we will present an algorithm such that the computational burden is independent of the size of the search grid.

5.1.3.1 Minimum Lower and Maximum Upper Bounds

For the BW approach, the starting point of the search grid for both the lower and upper bound is $1/N$, i.e., equal weights, see again Equations (5.2) and (5.3). For the end point, recall from Section 5.1.1, that as the lower (upper) bound goes towards negative (positive) infinity at some point it is sufficiently small (large) that it does not constrain the solution space anymore. Accordingly, there is a lower bound $\underline{\omega}^*$ at which the solution will not change if the lower bound is further decreased, i.e., $\underline{\omega} < \underline{\omega}^*$. This holds equivalently for the upper bound for $\overline{\omega} > \overline{\omega}^*$. We will refer to those values as the *minimum lower bound* $\underline{\omega}^*$ and *maximum upper bound* $\overline{\omega}^*$.[36] To evaluate candidate values for the lower and upper bound efficiently we do not want to include any lower bounds smaller (larger) than the minimum lower (maximum upper) bound.

At first glance the smallest and largest weights of the optimal weights solution (OW) could be the minimum lower and maximum upper. However, this is only true in specific scenarios. Recall that Table 5.1 presented the feasible values for the lower and upper

[36] To be clear, we neither try to estimate the minimum lower and maximum upper bound nor derive a general formula for them. We consider them solely for the purpose of defining the search grid for a given data set.

bound as well as the benchmark methods that are part of the solution space. Although looking at Table 5.1 might give the impression that the minimum lower and maximum upper bound, $\underline{\omega}^*$ and $\overline{\omega}^*$, are constant. i.e., independent of the corresponding other bound, this is not the case. There is not one minimum lower and maximum upper bound. On the contrary, the minimum lower bound $\underline{\omega}^*$ depends on the corresponding current value of the upper bound $\overline{\omega}$. Respectively, the maximum upper bound $\overline{\omega}^*$ is connected to the currently considered lower bound $\underline{\omega}$. Henceforth, both are functions of the corresponding other bound, i.e.,

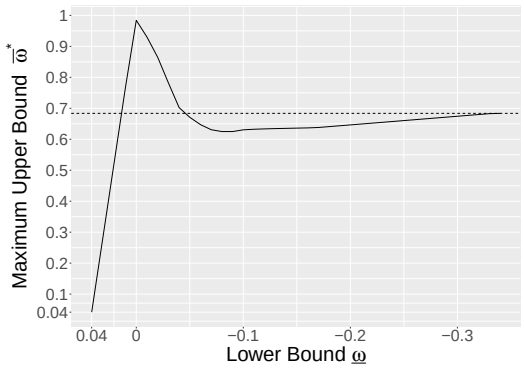
$$\text{Minimum lower bound: } \underline{\omega}^*(\overline{\omega}), \tag{5.4}$$

$$\text{Maximum upper bound: } \overline{\omega}^*(\underline{\omega}). \tag{5.5}$$

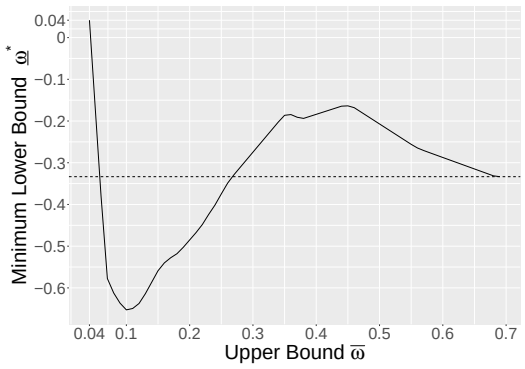
The fact that the maximum upper bound $\overline{\omega}^*$ depends on the current value of the lower bound $\underline{\omega}$ becomes intuitively clear when examining potential solutions of the PW and OW approaches. Both are located on the right side of Table 5.1, which represents solutions without an upper bound, i.e., $\overline{\omega} \to \infty$. The largest possible weight the PW approach can assign to any forecast is one, i.e., the maximum upper bound $\overline{\omega}^*(\underline{\omega} = 0) \leq 1$. Any weight larger than one would violate the non-negativity constraint ($\underline{\omega} = 0$) in conjunction with the unity constraint of the BW approach in Equation (5.1). In contrast, the OW solution ($\overline{\omega} \to \infty$ and $\underline{\omega} \to \infty$) can have weights larger one. Consequently, the maximum upper bound of the OW approach can be larger than one, i.e., $\overline{\omega}^*(\underline{\omega} \to -\infty) \geq 1$. Hence, the maximum upper bound of the PW and OW can be different. Similarly, an example that the minimum lower bound $\underline{\omega}^*$ depends on the currently considered upper bound $\overline{\omega}$ are the EW and OW solution. For the first, the upper bound is $1/N$ and, due to the unity constraint, the minimum lower bound $1/N$. Any lower bound smaller that, does not change the solution. For the OW approach ($\overline{\omega} \to \infty$) weights can be smaller $1/N$. Accordingly, the minimum lower bound of the OW can be smaller than the minimum lower bound of the EW approach. Based on those examples it becomes clear that the maximum upper and minimum lower bound can vary depending on the currently considered other bound.[37]

Beside the fact that the minimum lower and maximum upper bound are not constant, they also are not non-increasing or non-decreasing respectively. Figure 5.4 depicts the maximum upper and minimum lower bound for a simulated data set. It was generated by the simulation study presented in Section 3.2. We used a high correlation between forecast errors, i.e., $CM1$, a relative group distance $z = 0.5$ and the first group is significantly better than all others. The number of forecasts is $N = 24$. Figure 5.4(a) depicts the maximum upper bound on the ordinate and the abscissa shows the corresponding

[37] There are of course exceptions to it. If the OW solution does not contain weights greater one, i.e., no negative weights are present, the maximum upper and minimum lower bound of the PW and OW approach can be identical.



(a) Maximum upper bound for a given lower bound: $\overline{\omega}^*(\underline{\omega})$.



(b) Minimum lower bound for a given upper bound: $\underline{\omega}^*(\overline{\omega})$.

Figure 5.4. Minimum lower and Maximum upper bound (black). The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. We used a different setup for illustration purposes.

lower bound, again depicted in reverse. For Figure 5.4(b) the minimum lower bound is depicted by the ordinate while the abscissa shows the upper bound. In Figure 5.4(a) the horizontal black dashed line represents the largest weight of the OW approach, i.e., maximum upper bound if $\underline{\omega} \leq \underline{\omega}^*$. Similarly, in Figure 5.4(b) it is the smallest OW weight or minimum lower bound for $\overline{\omega} \geq \overline{\omega}^*$. The maximum upper bounds in Figure 5.4(a) starts at $1/N$. As the lower bound gets smaller, the maximum upper bound increases and becomes larger than the maximum upper bound of the OW approach (dashed line). Thereafter, it decreases, becomes again smaller than the maximum upper bound of the OW approach before transitioning to it. Similarly, the minimum lower bound in Figure 5.4(b) is even smaller than the minimum lower bound of the OW solutions for some upper bounds (e.g., $\overline{\omega} = 0.1$) and larger for others (e.g., $\overline{\omega} = 0.4$).

In summary, for any given lower bound $\underline{\omega}$ the maximum upper bound $\overline{\omega}^*(\underline{\omega})$ can be larger, smaller, or equal to the maximum upper bound for $\overline{\omega}^*(\underline{\omega} + \zeta)$ with $\zeta \in \mathbb{R}$. This holds similarly for the minimum lower bound. As a result, we cannot use the smallest and largest weights of the optimal weights solution to determine the most extreme $\overline{\omega}^*$ and $\underline{\omega}^*$. Instead, we have to determine them individually for each lower and upper bound respectively.

Lowermost and Uppermost Bound We can define a *lowermost bound* and an *uppermost bound*. Those are the most extreme values the minimum lower bound $\underline{\omega}^*(\overline{\omega})$ and maximum upper bound $\overline{\omega}^*(\underline{\omega})$ can have for a given lower and upper bound respectively. They are true in general for any pool of forecasts.

The maximum upper bound is identical to the largest weight of a solution given a certain lower bound. To find the uppermost bound, we need to determine the largest

possible weight a feasible solution can have given a lower bound. It is given if all weights but one are equal to the lower bound. To fulfill the unity constraint the last remaining weight has to offset all other weights. Accordingly, the uppermost bound is given if we assume that $w_j > w_i \; \forall \; i \in \{1, \ldots, N\}$ and $w_i = \underline{\omega} \; \forall \; i \in \{1, \ldots, N\} \setminus j$ with $\underline{\omega} \in (-\infty, 1/N]$. Then the sum of identical weights is

$$\sum_{\substack{i=1 \\ i \neq j}}^{N} w_i = (N-1)\underline{\omega}. \tag{5.6}$$

Accordingly, due to the unity constraint, for the maximum upper bound of any given pool of forecasts it holds that

$$\overline{\omega}^*(\underline{\omega}) \leq 1 - (N-1)\underline{\omega}. \tag{5.7}$$

If $\underline{\omega} = 0$ the uppermost bound is naturally one. If instead $\underline{\omega} < 0$, the second part of the equation $(N-1)\underline{\omega}$ is strictly negative and, thus, $1 - (N-1)\underline{\omega}$ is greater one.[38] If the lower bound $\underline{\omega}$ is greater one, the uppermost bound is smaller one accordingly.

The lowermost bound can be derived similarly by assuming $w_j < w_i \forall \; i \in \{1, \ldots, N\}$ while, $w_i = \overline{\omega} \; \forall \; i \in \{1, \ldots, N\} \setminus j$ with $\overline{\omega} \in [1/N, \infty)$. For the minimum lower bound it holds that

$$\underline{\omega}^*(\overline{\omega}) \geq 1 - (N-1)\overline{\omega}. \tag{5.8}$$

The uppermost and lowermost bounds of Equations (5.7) and (5.8) give an indication which values for the lower and upper bound have to be considered within a search grid. The problem is that the lowermost and uppermost bound depend on a value for the upper and lower bound, respectively. They only provide information for the most extreme minimum lower and maximum upper bound *conditional* on the other bound. For example, for $N = 4$ we can determine the lowermost bound for values for the lower bound, $\overline{\omega} = \frac{1}{N}, 0.1, 0.2, \ldots$. However, we still do not have an end point for $\overline{\omega}$. This holds accordingly for the uppermost / maximum upper bound which is conditional on $\underline{\omega}$. Hence, we can not use the lowermost and uppermost bound to determine the end point of the search grid.

The takeaway from the lowermost and uppermost bound is more computationally relevant. For a given search grid, it gives us information up to which value we actually have to calculate weights for pairs of candidate lower and upper bounds. For example, assume that there is a predefined search grid. With $N = 24$ for $\underline{\omega} = -0.1$ or -0.5 the uppermost bound is 3.3 and 12.5 respectively. As a result, all candidate values of $\overline{\omega} > 3.3$ where $\underline{\omega} = -0.1$ have the same weights and thus do not need to be estimated.

[38] The case of $N = 1$ is excluded as forecast combination with one forecast is trivial.

However, for $\underline{\omega} = -0.5$ candidate values of the upper bound do not have the same weights until $\overline{\omega} > 12.5$. As a result, we did not reduce the number of candidate pairs to evaluate, but the number of candidate pairs estimate weights for.

Unfortunately, this still results in unnecessary consumption of computational resources. Let us for example consider a lower bound $\underline{\omega} = -1$. Based on Equation (5.7), the uppermost bound is identical to N. For $N = 12$ and 24 we have to consider all upper bounds up to 12 and 24 respectively. If instead the lower bound is -4 or -8, the corresponding uppermost bound for $N = 12$ is 45 and 93, while for $N = 24$ we need to consider upper bounds up to 89 or 185, respectively. In comparison with some actual maximum upper (or minimum lower) bounds it becomes clear that the uppermost bound can be quite lose. For example, considering again Figure 5.4(a), for $\underline{\omega} = -0.2$ with $N = 24$. The uppermost bound is 5.6 while the actual maximum upper bound is roughly 0.65. Consequently, there is a lot more potential to reduce the number of candidate pairs for which weights have to be estimated by using the actual maximum upper and minimum lower bounds.

5.1.3.2 An Algorithm to Efficiently Evaluate Candidate Values of Bounds

Recall that the maximum upper (minimum lower) bounds for certain lower and upper bounds can be larger, smaller, or equal than the maximum upper (minimum lower) bound of the OW approach (see Figure 5.4). However, for a given pool of forecasts, the maximum upper bound $\overline{\omega}^*(\underline{\omega})$ does not change anymore for any lower bound $\underline{\omega} \leq min(\boldsymbol{\omega}^{OW})$. This is due to the fact that then neither the lower nor upper bound constraint the solution space anymore, i.e., the BW approach is identical to the OW approach. This holds accordingly for the minimum lower bound.

Based on this, we can create a simple and efficient algorithm to evaluate all candidate pairs of lower and upper bounds from a given search grid. Figure 5.5 illustrates which candidate pairs have to be evaluated, and it will be used to explain the procedure of the algorithm. It is designed similarly to Table 5.1 which illustrated the feasible bounds and benchmark methods. The abscissa in Figure 5.5 is located on top and shows the feasible values of the upper bound $\overline{\omega}$ and simultaneously the maximum upper bound $\overline{\omega}^*(\underline{\omega})$. The ordinate depicts the lower bound $\underline{\omega}$ as well as the minimum lower bound $\underline{\omega}^*(\overline{\omega})$.

The Algorithm

I. Calculate the smallest and largest weight of the optimal weights solution which we denote by $\underline{\omega}^{OW}$ and $\overline{\omega}^{OW}$. These weights are the minimum lower and maximum upper bound given the corresponding other weight as a bound, i.e., $\overline{\omega}^*(\underline{\omega}^{OW}) = \overline{\omega}^{OW}$ and vice versa. The dotted lines in Figure 5.5 show the position of both $\underline{\omega}^{OW}$

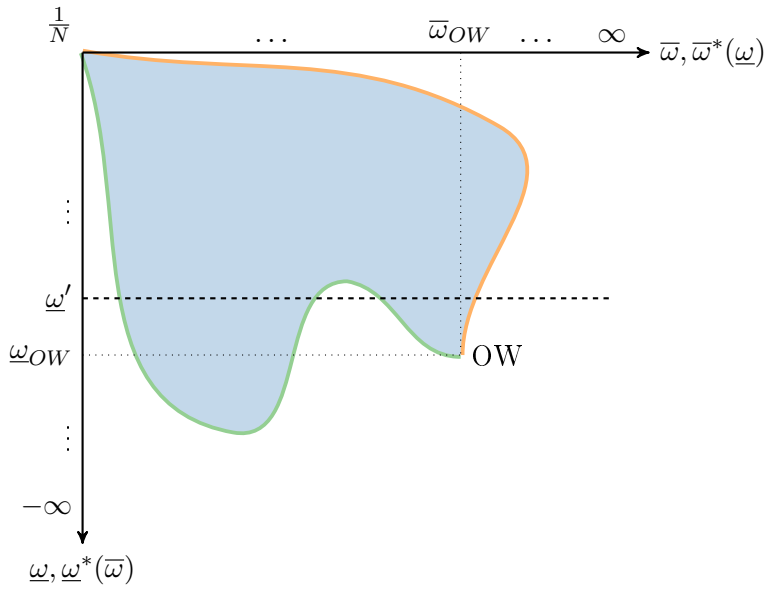


Figure 5.5. Illustration of the candidate pairs of lower and upper bound that need to be evaluated.

and $\overline{\omega}^{OW}$. If both values are used as the lower and upper bound, the estimated weights are identical to the OW approach, see the intersection of dotted lines.

II. In this step we calculate the maximum upper bounds depicted by the orange curve in Figure 5.5. For all $\underline{\omega} \geq \underline{\omega}^{OW}$ (likely negative values) calculate the weights from the BW approach without an upper bound, i.e., $\overline{\omega} \to \infty$. The largest weights of each solution is the maximum upper bound given the corresponding lower bound, i.e., $\overline{\omega}^*(\underline{\omega})$. Note that for $\underline{\omega} < \underline{\omega}^*_{OW}$ the maximum upper bound $\overline{\omega}^*(\underline{\omega})$ (orange curve) is identical to the maximum upper bound of OW, i.e., $\overline{\omega}^*(\underline{\omega}^{OW})$.

III. Similar to step II, we calculate the minimum lower bounds depicted by the green curve in Figure 5.5. For all $\overline{\omega} \leq \overline{\omega}^{OW}$ calculate the weights from the BW approach without a lower bound, i.e., $\underline{\omega} \to -\infty$. The smallest weights of each solution is the minimum lower bound given the corresponding upper bound, i.e., $\underline{\omega}^*(\overline{\omega})$. Note that for $\overline{\omega} > \overline{\omega}^*_{OW}$ the minimum lower bound $\underline{\omega}^*(\overline{\omega})$ (green curve) is identical to the minimum lower bound of OW, i.e., $\underline{\omega}^*(\overline{\omega}^{OW})$.

IV. With steps II and III we calculated the values of the green and orange curve, i.e., the minimum lower bounds $\underline{\omega}^*(\overline{\omega})$ and maximum upper bounds $\overline{\omega}^*(\underline{\omega})$. For all $\overline{\omega} = 1/N, \ldots, \overline{\omega}^{OW}$ with $\underline{\omega} \geq min(\underline{\omega}^*(\overline{\omega}))$ calculate the weights of candidate pair of lower and upper bounds if and only if $\overline{\omega} \leq \overline{\omega}^*(\underline{\omega})$ and $\underline{\omega} \geq \underline{\omega}^*(\overline{\omega})$.

In other words, the procedure is repeated for every lower bound that is larger than the smallest minimum lower bound and upper bound that is, simultaneously smaller than its maximum upper bound. With respect to figure Figure 5.5 we

consider every lower bound above the minimum of the green curve. To illustrate this step consider a specific value for the lower bound $\underline{\omega}'$ shown in Figure 5.5 with a black dashed line. For $\underline{\omega}'$ we only want to evaluate the upper bounds $\overline{\omega}$ that are smaller than the corresponding maximum upper bound, i.e., values left of the orange line. Every upper bound on the right of the orange line has the same weights for $\underline{\omega}'$. Similarly, we only need to calculate weights for the upper bounds for which the minimum lower bound is larger than $\underline{\omega}'$. Lower bounds below the green line do not result in other weights for the corresponding upper bounds. In Figure 5.5 the relevant candidate pairs are depicted by the blue area. For $\underline{\omega}'$ we only need to consider upper bounds where the dashed line coincides with the blue shaded area.

V. For all candidate pairs with both $\underline{\omega} \leq \underline{\omega}^*(\overline{\omega})$ and $\overline{\omega} \leq \overline{\omega}^*_{OW}$ weights are identical to $\underline{\omega}^*(\overline{\omega})$. In Figure 5.5 this concerns all candidate pairs that are both left of $\overline{\omega}^*_{OW}$ and below the green curve.

VI. For all candidate pairs with both $\overline{\omega} \geq \overline{\omega}^*(\underline{\omega})$ and $\underline{\omega} \geq \underline{\omega}^*_{OW}$ weights are identical to $\overline{\omega}^*(\underline{\omega})$. In Figure 5.5 this concerns all candidate pairs that are both above $\underline{\omega}^*_{OW}$ and right from the orange curve.

VII. For all candidate pairs with both $\overline{\omega} \geq \overline{\omega}^*_{OW}$ and $\underline{\omega} \leq \underline{\omega}^*_{OW}$ weights are identical to the OW solution. In Figure 5.5 this concerns all candidate pairs that are both right of $\overline{\omega}^*_{OW}$ and below $\underline{\omega}^*_{OW}$.

Applying this algorithm reduces the number of candidate values to be evaluated to its minimum. However, the search grid needs to be sufficiently large. This means that the largest candidate upper bound is greater than the largest maximum upper bound. This holds similarly for the smallest candidate lower bound. In terms of Figure 5.5, the largest (smallest) candidate upper (lower) bound needs to be right (below) of the orange (green) curve. However, one can simply use a very large (small) value for the candidate upper (lower) bound. If the search grid includes the smallest minimum lower bound and largest maximum upper bound, the algorithms computational time is independent of the number of candidate values.

5.2 Bounded Prior Weights

In Chapter 4 we used L_1 constraints to shrink weights towards either a fixed value κ or a prior weight vector $\dot{\boldsymbol{\omega}}$. Basically, the BW approach already includes solutions that result from imposing bounds around a fixed value κ. For $\kappa = 1/N$, all lower and upper bound basically can be interpreted as imposing bounds around this fixed value. This holds similarly for $\kappa = 0$ with $\underline{\omega} < 0$ and any $\overline{\omega}$. Recall that shrinkage towards a fixed value κ is a special case of shrinkage towards prior weights, if the prior weights are all

identical. However, the way we implemented the bounded weights approach, it does not incorporate shrinkage or bounds around prior weights. Therefore, we extend our approach of bounded weights to incorporated prior weights. The optimization problem is very similar to the bounded weights optimization problem of Equation (5.1). The only thing that changes is how we define the lower and upper bounds. The new optimization problem for our approach of *bounded prior weights* (BW($\dot{\omega}$)) is given by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}' \widehat{\boldsymbol{\Sigma}} \boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}' \mathbf{1} = 1, \\ & \omega_i \geq \underline{\omega}_i \qquad \forall\, i = 1, \dots, N, \\ & \omega_i \leq \overline{\omega}_i \qquad \forall\, i = 1, \dots, N \end{aligned} \tag{5.9}$$

In fact, there are no longer common lower and upper bound for all weights, but individual lower and upper bounds for each weight. Accordingly, there are $2N$ different bounds that we have to define which is a problem if we want to determine them by cross-validation. If we have the same number of candidate values for each individual lower and upper bound, the number candidate tuples or combination of hyperparameter values can easily become too large to determine trough cross-validation. If we consider P candidate values for each individual bound, the number of candidate tuples goes from P^2 in the BW approach to $P^{(2N)}$. In case of $N = 24$ forecasts and $P = 100$ candidate values, we would have to check 10^{96} candidate tuples which is multiple orders of magnitude larger than the current estimate for the number of atoms in the observable universe: 10^{81} (Tyson, Strauss, & Gott, 2016, pp. 19-20). Accordingly, we determine the individual lower and upper bounds by common deviations from the prior weights, i.e.,

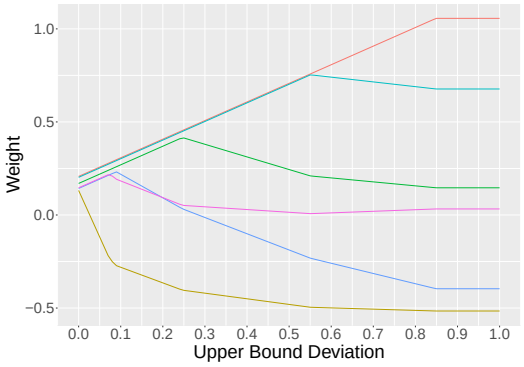
$$\underline{\omega}_i = \dot{\omega}_i - \underline{\beth} \qquad \forall i = 1, \dots, N, \tag{5.10}$$

$$\overline{\omega}_i = \dot{\omega}_i + \overline{\beth} \qquad \forall i = 1, \dots, N. \tag{5.11}$$

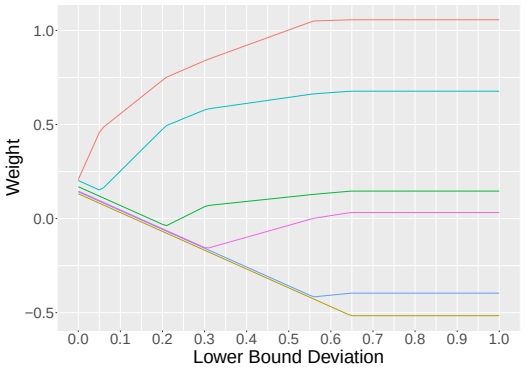
As a result, we again have two hyperparameters, the lower bound deviation $\underline{\beth}$ and upper bound deviation $\overline{\beth}$.[39] For both its holds that $\underline{\beth}, \overline{\beth} \geq 0$. By that they define an interval around the prior weights.

Beside the two deviations from the bounds, the prior weights we want to use have to be determined. Within this thesis we assume that the prior weights are a feasible solution to the forecast combination problem, i.e., it fulfills the unity constraint. As a consequence, for both the lower and upper bound deviation it holds that $\underline{\beth}, \overline{\beth} \in [0, \infty)$. If either the lower or upper bound deviation is zero, the solution is always equal to the prior weights. Even if the corresponding other bound deviation is larger, no weight can

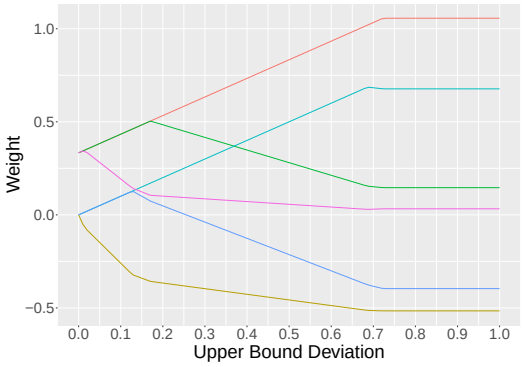
[39] Bet $\beth$ is a letter in the Hebrew alphabet.



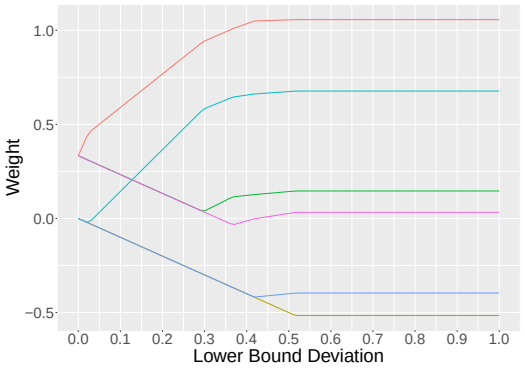
(a) Upper bound deviation $\overline{\sqsupseteq}$ with $\dot{\boldsymbol{\omega}} = \boldsymbol{\omega}^{IL}$.



(b) Lower bound deviation $\underline{\sqsupseteq}$ with $\dot{\boldsymbol{\omega}} = \boldsymbol{\omega}^{IL}$.



(c) Upper bound deviation $\overline{\sqsupseteq}$ with $\dot{\boldsymbol{\omega}} = \boldsymbol{\omega}^{E2}$.



(d) Lower bound deviation $\underline{\sqsupseteq}$ with $\dot{\boldsymbol{\omega}} = \boldsymbol{\omega}^{E2}$.

Figure 5.6. Illustration of forecast weight paths for Bounded Prior Weights with $\boldsymbol{\omega}^{IL}$ and $\boldsymbol{\omega}^{E2}$. Either the lower bound, Figures 5.6(a) and 5.6(c), or upper bound, Figures 5.6(b) and 5.6(d) is omitted. The data was created on the basis of the simulation study of Section 3.2 for $N = 24$ with $CM1$, an error variance similarity $z = 0.5$ and no special group. Six forecasts where chosen randomly out of 24 once and are used throughout this thesis.

change because no other weight can offset its change such that the solution remains feasible. Furthermore, if both deviations $\underline{\sqsupseteq}, \overline{\sqsupseteq}$ are sufficiently large such that they do not constrain the solution space, the bounded prior weights solution is identical to the OW solution.

Figure 5.6 depicts the weights path exemplary for bounded prior weights. In contrast to previously considered weight paths for the bounded weights approach in Section 5.1.2, for the bounded prior weights approach the abscissa has to depict the common deviation from the prior weight instead of the bound itself. In Figures 5.6(a) and 5.6(b) the inverse-loss weights of Equation (4.16) are used as prior weights. For Figures 5.6(c) and 5.6(d) the new shrinkage direction conditional group equal weights (CGEW) with two groups proposed in Section 4.2.3 is used. Considering both Figures 5.6(a) and 5.6(b) one can see the resemblance in behavior of the weight paths to the $\overleftrightarrow{EO}$ and $\overleftrightarrow{EPO}$ weight paths from the bounded weight approach depicted in Figures 5.2(a) and 5.2(b). As the

value on the abscissa increases, a group of weights increases or decreases together. To offset this change, individual weights deviate from this group and get an individual weights, see for example the yellow, purple, and blue line in Figure 5.6(a) or red and teal in Figure 5.6(b). However, if prior weights are used, the forecast within this group do not necessarily have the same weight, but they still increase by the same amount $\overline{\sqsupset}$ or $\underline{\sqsupset}$ respectively.

Recall that the idea behind CGEW proposed in Section 4.2.3 was to shrink a predefined set of forecasts towards their corresponding (conditional) equal weights and another group towards zero. While we used an L_1 constraint in Chapter 4, here we use lower and upper bounds or common deviations from the prior weights to be more precise. Figure 5.6(c) (Figure 5.6(d)) depicts the weights paths for using $\boldsymbol{\omega}^{E2}$ as the prior weights when the lower bound (upper bound) is omitted. By that, the resemblance to the bounded weights approach is even lager, see again Figure 5.2(a) (Figure 5.2(b)). There are two groups instead of one that have identical weights at $\overline{\sqsupset} = 0$ ($\underline{\sqsupset} = 0$ respectively). For one group the weight is the conditional equal weights of that group and for the other it is zero. As $\overline{\sqsupset}$ ($\underline{\sqsupset}$) increases, the weights of both identical groups increase (decrease) simultaneously. To offset this change at least one weight, first yellow then purple (first red then teal), from either group decreases (increases).

The findings of Section 5.1.3 regarding minimum lower bound and maximum upper bound hold accordingly for the lower and upper bound deviation. However, note that the minimum lower bound deviation is given by $\underline{\sqsupset}^*(\overline{\sqsupset}) = |min(\boldsymbol{\omega}^{OW} - \dot{\boldsymbol{\omega}})|$ and the maximum upper bound deviation is $\overline{\sqsupset}^*(\underline{\sqsupset}) = max(\boldsymbol{\omega}^{OW} - \dot{\boldsymbol{\omega}})$. Based on this, the algorithm for efficiently searching through pairs of candidate values of Section 5.1.3.2 can be easily adapted for bounded prior weights.

In this thesis we will use $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ from Section 4.2.3 as well as the inverse-loss weights $\boldsymbol{\omega}^{IL}$ of Equation (4.16) as shrinkage directions. Henceforth, we will refer to the general idea of the presented methods as Bounded Weights (BW). This includes Bounded Prior Weights. If we use the Bounded Weights approach without prior weights we refer to it as BW($\cdot$). If prior weights are used, we denote that by BW($\dot{\omega}$) and if we refer to specific prior weights we replace $\dot{\omega}$ with the corresponding prior weight, e.g., BW($\boldsymbol{\omega}^{E2}$).

5.3 Application: Simulation Study

In this section we analyze and compare the forecast accuracy of forecast combination with bounded (prior) weights. To this end, we use the simulation study introduced in Section 3.2. For a brief summary of the designed scenarios see Section 3.2.3 and the used forecast error correlation matrices are depicted in Section 3.2.2.2. Note that the following analysis is conducted in the same way as Section 4.3, i.e., oftentimes we use the same tools and figures to analyze the results. Therefore, the reader is refereed to the simulation study in Chapter 4 for a more detailed overview as well as introduction into tables and figures. Similar to Section 4.3 we first consider an ex post analysis in Section 5.3.1, i.e., we choose the single best lower and upper bound (or common deviation in case of bounded prior weights) for each test set as a whole ex post. After that in Section 5.3.2 we will analyze the out-of-sample forecasting performance of the bounded weights methods. To this end, we estimate the hyperparameters for each observation in the test set by cross-validation. Again, out-of-sample forecasts are based on a rolling window of size 40, and we use the last ten of those observations for cross-validation.

For both analysis we need to define which candidate values are considered for the hyperparameters. As we discussed before, the number of candidate pairs is much larger for the bounded (prior) weights approaches then it is for the L_1 methods. Recall that we can simply choose a very larger (small) value for the highest (smallest) lower and upper bound or deviation respectively. However, within a simulation study, an extensive empirical analysis or in a business context where one needs daily forecasts for multitude of time series it can get too computationally demanding. Additionally, statistical software can get slow because the resulting objects with weights and forecasts become too large if the largest grid value is too large. Only the former problem can be solved by the algorithm for an efficient hyperparameter search that we introduced in Section 5.1.3.2. Accordingly, we have to choose a reasonable larger (small) value for the search grids.

Search Grid for the Bounded Weights Methods Within this simulation study we will consider lower bounds ($\underline{\omega}$) down to -10, upper bound ($\overline{\omega}$) up to 10 (BW($\cdot$)) as well as lower and upper bound deviations ($\underline{\dot{\omega}}, \overline{\dot{\omega}}$) of 10 (BW($\dot{\omega}$)). Note that we chose a smaller value compared to the L_1 methods (50), because the bounds hold for each weight separately. For the L_1 method the hyperparameter determines the overall absolute deviation of all weight from the shrinkage directions.

In Section 5.1 we extended feasible lower bound values for both LB and BW($\cdot$) to also include positive values. This corresponds to a minimum contribution of each forecast. Accordingly, the grid for BW($\cdot$) starts for both the lower and upper bound at equal weights, i.e., $1/24 = 0.04167$. We use 0.1 increments for the candidate values of the bounds, as we did for the L_1 methods in Section 4.3. However, if we round $1/24$ down

to the first decimal place, the result is zero. By that we would completely neglect the transition from EW to PW that is imposed by the lower bound, i.e., $\overleftrightarrow{EP}$. Therefore, we use increments of 0.01 for the lower bound between $1/24$ and 0. The search grid for the lower bound is

$$(1/24, 0.04, 0.03, \ldots, 0, -0.1, -0.2, \ldots, -10), \tag{5.12}$$

and for the upper bound it is

$$(1/24, 0.1, 0.2, \ldots, 10). \tag{5.13}$$

In case of the bounded prior weights approaches it is more straightforward. The grid of candidate values for both the lower and upper deviation start at zero and increases to 10.[40] Note that we defined the search grid prior to the analysis.

5.3.1 Ex Post Analysis: Bounded Weights

In this section present the overall result in Section 5.3.1.1 and with respect to the error correlations, error variance similarity and special groups in Section 5.3.1.2.

5.3.1.1 Ex Post Analysis: Overall Results

Both Tables 5.3 and 5.4 show the average MSE values over all test sets for each scenario. Both for the benchmarks and Bounded Weights methods. It is designed in the same way as Tables 4.1 and 4.2. Note that we use Bounded Weights methods (BW) if we refer to all BW($\cdot$), BW($\dot{\omega}$) and LB together. At first glance, one can see that Bounded Weights oftentimes have smaller MSE values than the benchmark methods, in particular BW($\cdot$).

To further analyze the Bounded Weights methods we again take a look at the percentages of how often each method has the smallest MSE (possibly among multiple methods with the same MSE), the average rank and distances in the same way we did in Table 4.3. Note that a comparison of average ranks between Chapter 4 and Chapter 5 is not sensible, because they are calculated only based on the methods discussed in the respective chapter.[41]

The results from Table 5.5 show that the Bounded Weights methods are preferable to the benchmarks with respect to percentage of being best, average rank and distance as well as distribution of rank across the scenarios, see also Figure 5.7. Moreover, the results clearly show that the BW($\cdot$) method is by far superior not only compared to the benchmarks but also to BW($\dot{\omega}$) as well as LB. It has the smallest MSE for almost three-quarter of all scenarios, i.e., 53 in total. Moreover, it is the most consistent

[40] Recall that both the lower bound deviation and upper bound deviation are defined to be non-negative, see Equations (5.10) and (5.11).

[41] We compare all methods in an empirical analysis in Chapter 7.

CM	z	SG	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
1	0.05	none	0.98	0.97	2.11	0.98	**0.89**	0.90	0.90	**0.89**	0.93
		first	0.97	0.93	1.89	0.96	**0.84**	**0.84**	**0.84**	**0.84**	0.90
		last	1.01	0.99	2.04	1.00	**0.89**	**0.89**	**0.89**	**0.89**	0.95
		both	0.97	0.93	1.75	0.96	**0.80**	**0.80**	**0.80**	**0.80**	0.87
1	0.20	none	1.16	0.92	1.03	1.12	0.52	**0.51**	**0.51**	0.52	0.59
		first	1.12	0.73	0.52	1.04	**0.25**	**0.25**	**0.25**	**0.25**	0.28
		last	1.24	0.95	0.84	1.18	0.42	0.42	**0.41**	0.42	0.49
		both	1.15	0.72	0.45	1.05	0.23	0.23	**0.22**	0.23	0.25
1	0.50	none	1.53	0.80	0.42	1.33	**0.21**	**0.21**	**0.21**	0.22	0.24
		first	1.38	0.30	0.09	0.85	0.05	0.05	0.05	**0.04**	0.05
		last	1.64	0.82	0.35	1.38	**0.18**	**0.18**	**0.18**	0.19	0.21
		both	1.46	0.30	0.08	0.85	0.05	**0.04**	**0.04**	**0.04**	0.05
2	0.05	none	0.56	0.62	1.37	0.56	**0.54**	0.57	0.57	0.56	0.55
		first	0.56	0.62	1.34	0.56	**0.54**	0.57	0.56	0.55	0.55
		last	0.57	0.64	1.37	0.57	**0.55**	0.58	0.57	0.57	0.56
		both	0.57	0.63	1.35	0.57	**0.55**	0.58	0.57	0.57	0.56
2	0.20	none	0.68	0.68	1.42	0.66	**0.61**	0.62	0.62	0.63	0.64
		first	0.64	0.56	1.11	0.60	**0.52**	0.53	0.53	0.53	0.54
		last	0.68	0.66	1.31	0.65	**0.59**	**0.59**	**0.59**	0.61	0.63
		both	0.66	0.54	1.00	0.60	**0.49**	0.50	0.50	0.50	0.53
2	0.50	none	0.87	0.64	1.10	0.76	0.57	0.57	0.57	**0.56**	0.61
		first	0.80	0.26	0.35	0.49	0.22	0.21	0.21	**0.20**	0.23
		last	0.96	0.66	1.05	0.81	0.56	0.55	0.55	**0.54**	0.61
		both	0.83	0.26	0.33	0.49	0.21	**0.20**	**0.20**	**0.20**	0.23
3	0.05	none	**0.25**	0.32	0.63	0.26	**0.25**	0.28	0.27	0.26	**0.25**
		first	0.25	0.31	0.63	0.25	**0.24**	0.28	0.26	0.25	**0.24**
		last	**0.25**	0.32	0.63	**0.25**	**0.25**	0.28	0.27	**0.25**	**0.25**
		both	0.25	0.32	0.64	0.25	**0.24**	0.28	0.27	0.25	**0.24**
3	0.20	none	0.30	0.36	0.71	0.30	**0.29**	0.31	0.30	**0.29**	0.30
		first	0.29	0.32	0.63	**0.27**	**0.27**	0.28	**0.27**	**0.27**	0.28
		last	0.30	0.36	0.71	**0.29**	**0.29**	0.30	**0.29**	**0.29**	0.30
		both	0.30	0.32	0.61	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	0.28
3	0.50	none	0.40	0.39	0.75	0.35	0.35	0.35	**0.34**	0.35	0.36
		first	0.35	0.19	0.35	0.22	**0.18**	0.23	0.22	0.21	0.19
		last	0.43	0.40	0.76	0.36	0.36	0.35	**0.34**	0.36	0.38
		both	0.37	0.20	0.36	0.22	**0.19**	0.23	0.23	0.21	**0.19**

Table 5.3. Simulation study results of benchmark and Bounded Weights methods forecast combination methods for correlation matrices CM1, CM2 and CM3 (ex post analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

method. Its average rank is 2.01 and the next larger average rank is 2.97 for BW($\boldsymbol{\omega}^{IL}$).[42] Figure 5.7 shows that its rank is smaller or equal to three in 75% of the cases. Overall the interquartile range (IQR), i.e., the box of the boxplot, is noticeably smaller than the IQR of other methods. To put it differently, it usually has a better rank than other methods.

[42] Recall that if more than method have the same smallest MSE value, they are ranked in an arbitrary order and then the average rank is used for all of them. As a result although BW has the smallest MSE for about three-quarter of all scenarios, an average rank of two is plausible.

CM	z	SG	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
4	0.05	none	0.63	0.69	1.57	0.63	**0.62**	0.67	0.66	0.63	**0.62**
		first	0.65	0.70	1.61	0.65	**0.63**	0.68	0.67	0.64	0.64
		last	0.66	0.72	1.64	0.66	**0.64**	0.70	0.69	0.65	0.65
		both	0.64	0.68	1.59	0.63	**0.62**	0.66	0.65	0.63	**0.62**
4	0.20	none	0.79	0.78	1.78	0.76	**0.72**	0.75	0.74	0.74	0.74
		first	0.73	0.67	1.48	0.68	**0.62**	0.65	0.64	0.64	0.64
		last	0.80	0.78	1.74	0.77	**0.72**	0.74	0.74	0.74	0.75
		both	0.77	0.68	1.51	0.70	**0.64**	0.66	0.65	0.66	0.66
4	0.50	none	1.03	0.79	1.50	0.91	**0.73**	0.76	0.75	0.75	0.75
		first	0.90	0.30	0.45	0.56	**0.28**	0.30	0.30	0.30	**0.28**
		last	1.10	0.79	1.47	0.93	**0.74**	0.77	0.76	0.76	0.77
		both	1.02	0.31	0.46	0.60	**0.28**	0.30	0.30	0.30	**0.28**
5	0.05	none	0.44	0.40	0.84	0.45	**0.35**	0.42	0.42	0.40	0.38
		first	0.44	0.40	0.85	0.45	**0.35**	0.43	0.42	0.40	0.37
		last	0.44	0.41	0.88	0.46	**0.36**	0.44	0.43	0.41	0.38
		both	0.43	0.41	0.87	0.45	**0.35**	0.43	0.43	0.41	0.37
5	0.20	none	0.51	0.51	1.06	0.54	**0.44**	0.53	0.52	0.49	0.46
		first	0.48	0.49	0.98	0.50	**0.42**	0.50	0.49	0.47	0.44
		last	0.52	0.54	1.13	0.56	**0.47**	0.56	0.56	0.53	0.49
		both	0.49	0.53	1.06	0.51	**0.45**	0.52	0.52	0.49	0.47
5	0.50	none	0.67	0.66	1.21	0.69	**0.59**	0.66	0.65	0.63	0.62
		first	0.56	0.31	0.42	0.45	**0.29**	0.32	0.32	0.31	0.30
		last	0.66	0.67	1.22	0.68	**0.60**	0.66	0.66	0.64	0.62
		both	0.59	0.33	0.44	0.46	**0.31**	0.34	0.34	0.34	**0.31**
6	0.05	none	0.79	0.63	1.39	0.77	**0.57**	0.60	0.60	0.59	0.61
		first	0.78	0.60	1.32	0.75	**0.55**	0.57	0.57	0.56	0.58
		last	0.80	0.63	1.37	0.78	**0.57**	0.60	0.60	0.59	0.61
		both	0.80	0.61	1.30	0.77	**0.55**	0.57	0.57	0.56	0.60
6	0.20	none	0.97	0.64	1.16	0.89	0.53	0.53	0.53	**0.52**	0.61
		first	0.94	0.52	0.95	0.81	0.44	**0.43**	**0.43**	**0.43**	0.50
		last	1.02	0.64	0.96	0.91	0.47	**0.46**	**0.46**	**0.46**	0.56
		both	0.97	0.51	0.80	0.81	0.39	**0.38**	**0.38**	**0.38**	0.47
6	0.50	none	1.28	0.61	0.85	1.02	0.42	**0.41**	**0.41**	**0.41**	0.49
		first	1.22	0.26	0.42	0.62	0.21	**0.20**	**0.20**	**0.20**	0.23
		last	1.39	0.62	0.69	1.05	0.35	**0.34**	**0.34**	0.35	0.42
		both	1.31	0.26	0.38	0.62	0.19	0.19	**0.18**	0.19	0.22

Table 5.4. Simulation study results of benchmark and Bounded Weights methods forecast combination methods for correlation matrices CM4, CM5 and CM6 (ex post analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

BW($\dot{\omega}$) has best MSE between for a quarter to a third of scenarios. Using the inverse-loss average weights as prior weights has the smallest average rank of all considered prior weights and the smallest IQR. The LB approach has a noticeably smaller percentage of being the best method, in particular compared to BW(·). If we directly compare the BW(·) approach to LB, the former has on average a 5% smaller MSE. The largest improvement for BW(·) compared to LB is about 17%. However, based on this ex post analysis the BW(·) approach never has a larger MSE than the LB approach by design as BW(·) nests LB.

	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
Smallest MSE (%)	2.78	0.00	0.00	5.56	**73.61**	23.61	33.33	30.56	13.89
Avg Rank	7.16	6.30	8.34	6.47	**2.01**	4.37	3.65	2.97	3.74
Avg Distance	0.32	0.12	0.55	0.22	**0.00**	0.03	0.02	0.02	0.03

Table 5.5. Key figures for the MSE values of benchmark and Bounded Weights methods over all simulation study scenarios (ex post analysis). Smallest MSE (%) — Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank — Average rank of a method where a smaller rank is favorable. Avg Distance — Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

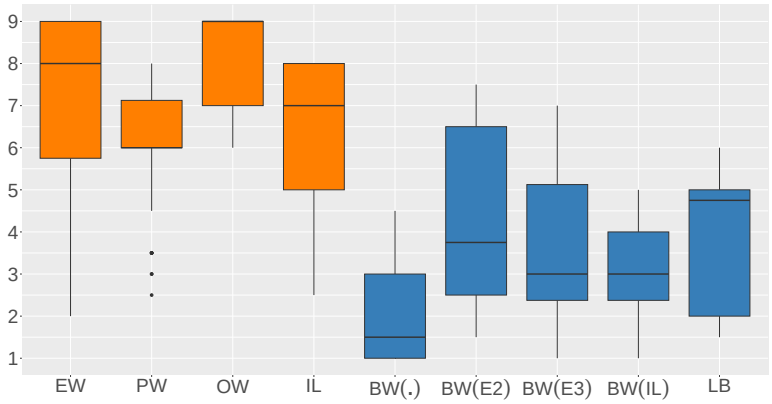


Figure 5.7. Boxplot of ranks across benchmarks and Bounded Weights methods for the ex post analysis.

Nevertheless, the presented results provide clear evidence that both using an upper bound and that using prior weights improves the out-of-sample forecast accuracy compared to the benchmarks and only using a lower bound.

We want to emphasize that we compare all methods against each other, i.e., also the different Bounded Weights methods. Due to that one may underrate our proposed Bounded Weights methods as a whole. Therefore, before we analyze the results note that for all 72 scenarios there is always a Bounded Weights methods that either has a smaller MSE than all benchmark methods or the same MSE. In fact for 67 scenarios the MSE of at least one Bounded Weights methods has a strictly smaller MSE than the benchmarks. Using the Bounded Weights methods is the superior strategy for the considered ex post analysis.

5.3.1.2 Ex Post Analysis: Groups of Scenarios

In what follows we analyze the performance of the forecast combination methods with respect to correlation matrices, error variance similarities and special groups. Table 5.6 shows how often each method has the smallest MSE for each correlation matrix Shrink-

	EW	PW	OW	IL	BW($\cdot$)	BW($\dot{\omega}$)			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
CM1	0.00	0.00	0.00	0.00	58.33	66.67	**83.33**	58.33	0.00
CM2	0.00	0.00	0.00	0.00	**66.67**	16.67	16.67	33.33	0.00
CM3	16.67	0.00	0.00	33.33	**83.33**	8.33	41.67	41.67	41.67
CM4	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	33.33
CM5	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	8.33
CM6	0.00	0.00	0.00	0.00	33.33	50.00	**58.33**	50.00	0.00

Table 5.6. Percentage of scenarios for which benchmarks and Bounded Weights methods have the smallest MSE with respect to the error correlation matrix (ex post analysis). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

age towards prior weights BW($\dot{\omega}$) is oftentimes the best method for highly correlated forecast errors (CM1 and CM6). Recall, that we have observed a similar result for the L_1 constraints from Section 4.3.1, see also Table 4.4. The BW($\cdot$) approach is more often superior as the error correlations decrease (CM2 with medium and CM3 with low correlations, 0.5 and 0.2 respectively) and for mixed correlation matrices (CM4 and CM5). For both CM4 and CM5 the BW($\cdot$) approach is always the best or one of the best methods. Using only a lower bound (LB) is the closest method to BW($\cdot$) for small error correlations (CM3). Recall that we extended the interval of feasible lower bounds compared to Radchenko et al. (2023) such that it also can have positive values. As a result, the lower bound can have values close to EW. CM3 has the smallest error correlations and thus is the closest to EW being optimal (no error correlation). The average lower bound used by LB across all scenario in CM3 is roughly $\underline{\omega} = 0.03$. This shows that the extension of the feasible lower bound interval we proposed improves forecast accuracy. For LB with $\underline{\omega} = 0.03$ and $N = 24$ the largest ever possible weight within a solution is 0.31. However, it is more likely that the largest weight is smaller as this is the most extreme case for which there is only one possible weight constellation. For the BW($\cdot$) approach the average lower and upper bound over all scenarios of CM3 are about 0.01 and 0.13. This provides clear evidence that to allow for a small deviation from equal weights and then estimating the weights within the forecast combination optimization problem provides huge benefits as one can see when comparing the results of EW and BW($\cdot$).

Figure 5.8 depicts the average ranks and distances sorted by the correlation matrices. In general, the ranks provide similar findings to the percentages of being the best

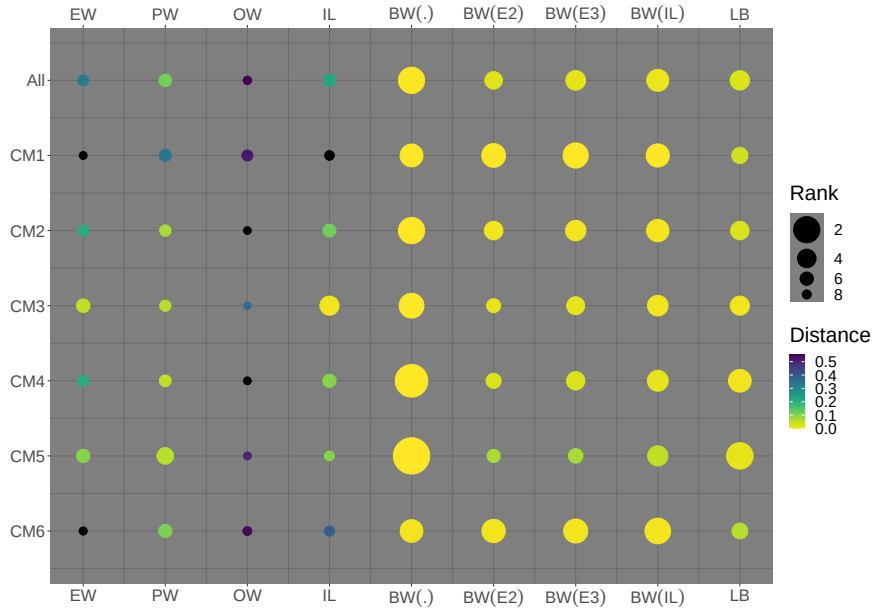


Figure 5.8. Illustration of average ranks and distances of the benchmarks and Bounded Weights for different correlations matrices (ex post analysis).

method. Overall the BW(·) approach has the smallest rank, i.e., it is the most consistent method. BW(·) and LB have a similar behavior with respect to the correlation matrices, i.e., they show a similar direction of improvement when comparing one to another correlation matrix. The bounded weights around prior weights are more consistent for highly correlated forecast errors, however, they only have a marginally smaller rank than BW(·). For CM5 the average distance of BW($\dot{\omega}$) to BW(·) (which has the smallest MSE for each scenario) is larger than for all other correlation matrices.

Table 5.7 presents how often each method has the smallest MSE with respect to the error variance similarities (rows one to three) and special groups (rows four to seven). In addition to that Figure 5.9 shows the average ranks and distance to the best method scenario-wise. BW(·) is always superior or as good as other methods if the error variance of forecast are similar in terms of percentage being the best method, average rank and distance. The more dissimilar the forecast error variances become, the less often it is the best method (Table 5.7). Nevertheless, no other method has the smallest MSE more often, higher rank or smaller distance (Figure 5.9). In general, the bounds around prior weights become more viable the more dissimilar forecast are. In those cases they are the best method more often and have better average ranks, in particular BW($\boldsymbol{\omega}^{E3}$), as one can see by the size of the circle in Figure 5.9.

With respect to special groups it appears that no method is in particular tailored to a specific special group of forecasts. Comparing the ranks and distances to *All* scenarios in Figure 5.9 reveals that there is not much of difference, i.e., similar circle size (rank) and color (distance). The only exception is BW($\dot{\omega}$) with $\boldsymbol{\omega}^{E2}$ which has a

	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
$z = 0.05$	8.33	0.00	0.00	4.17	**100.00**	12.50	12.50	20.83	25.00
$z = 0.20$	0.00	0.00	0.00	12.50	**70.83**	29.17	45.83	37.50	0.00
$z = 0.50$	0.00	0.00	0.00	0.00	**50.00**	29.17	41.67	33.33	16.67
none	5.56	0.00	0.00	0.00	**72.22**	16.67	22.22	27.78	11.11
first	0.00	0.00	0.00	5.56	**77.78**	22.22	27.78	38.89	11.11
last	5.56	0.00	0.00	11.11	**72.22**	27.78	44.44	27.78	5.56
both	0.00	0.00	0.00	5.56	**72.22**	27.78	38.89	27.78	27.78

Table 5.7. Percentage of scenarios for which the benchmarks and Bounded Weights methods have the smallest MSE with respect to the error variance similarity and special groups (ex post analysis). The total number of scenarios for each error variance similarity is 24 and for special groups it is 18. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

noticeable increase in percentage of being the best if the last group has a worse forecast performance (Table 5.7). Additionally, LB is more tailored towards the scenario where both a noticeably good and bad group are present. However, both considered methods are still inferior to BW(·).

In summary, the presented result provide evidence that for our simulation study without hyperparameter estimation, imposing both a lower and an upper bound is the dominant strategy among the Bounded Weights methods and benchmarks. Additionally, using prior weights is preferable compared to the benchmarks and LB. With respect to LB, allowing for positive lower bound can improve forecast accuracy.

5.3.2 Out-Of-Sample: Bounded Weights with Hyperparameter Estimation

In this section we analyze the out-of-sample forecast accuracy of the Bounded Weights methods (BW(·)). Note that we use the same time series as in Section 5.3.1. However, we look at the pseudo out-of-sample forecasting performance of the considered methods. Recall, there are two differences to the ex post analysis from Section 5.3.1. First, we determine the hyperparameter a priori based on only past information. Second, we re-estimate or -determine the hyperparameters for each observation in the test individually based on cross-validation, recall Section 2.1. Due to the latter change, the MSE of the considered methods can be smaller than the MSE of the ex post analysis. To put it differently, for the ex post analysis we chose one hyperparameter for the whole test set. If we instead use the best hyperparameters for each observation in the test set, the resulting MSE is at most equal to the MSE of a fixed hyperparameter over the whole test set. Accordingly, if we re-estimate the hyperparameter for each observation in the test set and this estimation is precise and accurate, we can achieve a smaller MSE.

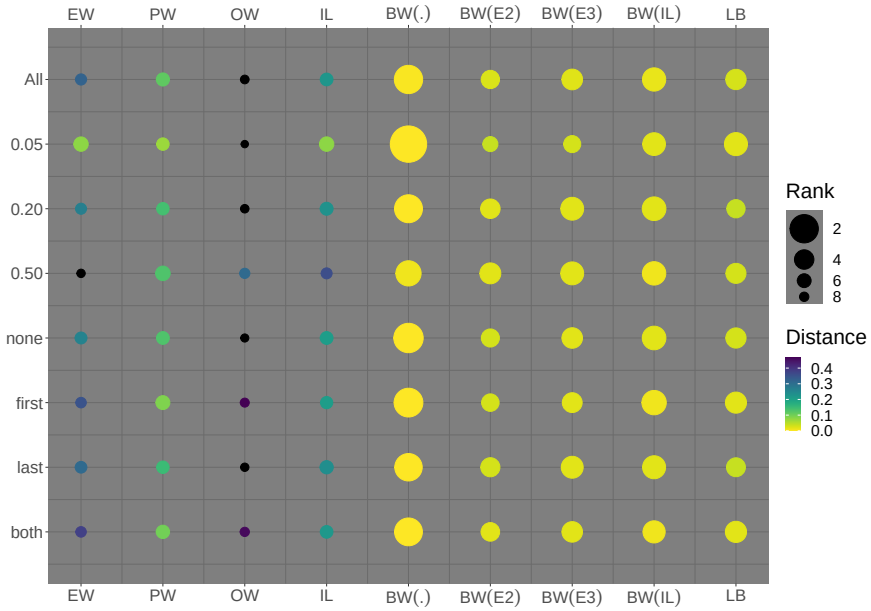


Figure 5.9. Illustration of average ranks and distances of the benchmarks and Bounded Weights methods for error variance similarities and special groups (ex post analysis).

5.3.2.1 Out-Of-Sample: Overall Results

Tables 5.8 and 5.9 show the average MSE across all 200 test sets for each scenario. Both Tables 5.8 and 5.9 follow the exact same structure as Tables 5.3 and 5.4 respectively. The results reveal that BW(·) oftentimes has a larger MSE than other methods. Accordingly, the unambiguous superiority of the BW(·) approach does not translate to pseudo out-of-sample forecasting. At least if hyperparameters are estimated based on the used cross-validation approach. Consider for example the scenarios 1/0.05/·. For all four special groups the PW method has the smallest MSE with a noticeable difference to the BW(·) MSE. This first result is similar to the one from Chapter 4 for the L_1 methods which showed superior forecasting performance in the ex post analysis that is not achieved if hyperparameters are estimated.

Table 5.10 shows how often each method has the smallest MSE, its average rank and distance to the best method scenario-wise. In addition to that, Figure 5.10 shows boxplots of the ranks over all 72 scenarios for each method. The PW approach has the smallest MSE most often, followed by the inverse-loss weights and equal weights. Based on Figure 5.10 one can see that for 70% of the scenarios PW has a rank smaller or equal to 5.5. In comparison, EW and IL have a larger IQR than PW, i.e., they are less consistent. Using both a lower and upper bound (BW) as well as only using a lower bound (LB) only has the smallest MSE in the same three scenarios (1/0.20/first, 1/0.20/both and 1/0.50/last). The average ranks of BW (4.49) and LB (4.14) are noticeably larger compared to PW (3.35) but they have a better rank than EW (5.42).

CM	z	SG	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
1	0.05	none	0.98	**0.97**	2.11	0.98	1.06	1.07	1.06	1.07	1.05
		first	0.97	**0.93**	1.89	0.96	1.00	1.00	1.00	1.01	1.01
		last	1.01	**0.99**	2.04	1.00	1.06	1.06	1.06	1.07	1.08
		both	0.97	**0.93**	1.75	0.96	0.97	0.96	0.96	0.97	0.99
1	0.20	none	1.16	0.92	1.03	1.12	0.66	**0.64**	**0.64**	0.65	0.70
		first	1.12	0.73	0.52	1.04	**0.33**	**0.33**	**0.33**	**0.33**	**0.33**
		last	1.24	0.95	0.84	1.18	0.54	**0.53**	**0.53**	0.54	0.59
		both	1.15	0.72	0.45	1.05	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**
1	0.50	none	1.53	0.80	0.42	1.33	0.27	**0.26**	0.27	0.27	0.27
		first	1.38	0.30	0.09	0.85	0.06	0.06	**0.05**	**0.05**	0.06
		last	1.64	0.82	0.35	1.38	**0.23**	**0.23**	**0.23**	0.24	**0.23**
		both	1.46	0.30	0.08	0.85	0.06	**0.05**	**0.05**	**0.05**	0.06
2	0.05	none	**0.56**	0.62	1.37	**0.56**	0.62	0.65	0.64	0.64	0.61
		first	**0.56**	0.62	1.34	**0.56**	0.63	0.66	0.65	0.64	0.61
		last	**0.57**	0.64	1.37	**0.57**	0.65	0.67	0.66	0.66	0.62
		both	**0.57**	0.63	1.35	**0.57**	0.64	0.66	0.65	0.65	0.63
2	0.20	none	0.68	0.68	1.42	**0.66**	0.73	0.73	0.73	0.74	0.71
		first	0.64	**0.56**	1.11	0.60	0.63	0.62	0.62	0.63	0.63
		last	0.68	0.66	1.31	**0.65**	0.71	0.70	0.69	0.72	0.71
		both	0.66	**0.54**	1.00	0.60	0.60	0.59	0.59	0.60	0.61
2	0.50	none	0.87	**0.64**	1.10	0.76	0.71	0.68	0.68	0.67	0.71
		first	0.80	0.26	0.35	0.49	0.27	0.26	**0.25**	**0.25**	0.26
		last	0.96	**0.66**	1.05	0.81	0.69	**0.66**	**0.66**	0.67	0.73
		both	0.83	0.26	0.33	0.49	0.27	**0.25**	**0.25**	**0.25**	0.26
3	0.05	none	**0.25**	0.32	0.63	0.26	0.29	0.32	0.31	0.29	0.28
		first	**0.25**	0.31	0.63	**0.25**	0.28	0.31	0.30	0.28	0.27
		last	**0.25**	0.32	0.63	**0.25**	0.29	0.32	0.31	0.29	0.28
		both	**0.25**	0.32	0.64	**0.25**	0.28	0.31	0.30	0.28	0.27
3	0.20	none	**0.30**	0.36	0.71	**0.30**	0.34	0.36	0.35	0.34	0.33
		first	0.29	0.32	0.63	**0.27**	0.32	0.32	0.31	0.31	0.31
		last	0.30	0.36	0.71	**0.29**	0.33	0.34	0.33	0.33	0.33
		both	0.30	0.32	0.61	**0.27**	0.32	0.31	0.30	0.31	0.31
3	0.50	none	0.40	0.39	0.75	**0.35**	0.41	0.40	0.39	0.40	0.41
		first	0.35	**0.19**	0.35	0.22	0.23	0.27	0.26	0.24	0.22
		last	0.43	0.40	0.76	**0.36**	0.43	0.40	0.39	0.41	0.43
		both	0.37	**0.20**	0.36	0.22	0.24	0.27	0.26	0.24	0.23

Table 5.8. Simulation study results of benchmark and Bounded Weights methods forecast combination methods for correlation matrices CM1, CM2 and CM3 (out-of-sample analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

Both BW(·) and LB are more consistent as one can see looking at the smaller IQR in Figure 5.10, i.e., they have oftentimes a higher rank but less variation in ranks. Nevertheless, as for BW(·) and LB the average rank is higher, this is not favorable. For example for BW(·) only in about 25% of scenarios the rank is smaller or equal to about 3.5 while PW has a rank of 1 for at least 25% of scenarios. However, the average distance of BW(·) and LB to the best method presented in Table 5.10 are the smallest among all other methods. Accordingly, they oftentimes are relatively close to the best method, although there are not the best method themselves.

CM	z	SG	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
4	0.05	none	**0.63**	0.69	1.57	**0.63**	0.73	0.78	0.77	0.74	0.70
		first	**0.65**	0.70	1.61	**0.65**	0.74	0.78	0.78	0.75	0.71
		last	**0.66**	0.72	1.64	**0.66**	0.75	0.81	0.79	0.76	0.72
		both	0.64	0.68	1.59	**0.63**	0.72	0.77	0.76	0.74	0.69
4	0.20	none	0.79	0.78	1.78	**0.76**	0.86	0.88	0.87	0.86	0.84
		first	0.73	**0.67**	1.48	0.68	0.74	0.75	0.74	0.75	0.73
		last	0.80	0.78	1.74	**0.77**	0.87	0.88	0.87	0.87	0.86
		both	0.77	**0.68**	1.51	0.70	0.77	0.77	0.76	0.77	0.75
4	0.50	none	1.03	**0.79**	1.50	0.91	0.90	0.91	0.90	0.92	0.88
		first	0.90	**0.30**	0.45	0.56	0.36	0.39	0.39	0.38	0.33
		last	1.10	**0.79**	1.47	0.93	0.91	0.91	0.91	0.92	0.88
		both	1.02	**0.31**	0.46	0.60	0.36	0.39	0.39	0.39	0.33
5	0.05	none	0.44	**0.40**	0.84	0.45	0.43	0.50	0.50	0.47	0.43
		first	0.44	**0.40**	0.85	0.45	0.42	0.51	0.50	0.47	0.42
		last	0.44	**0.41**	0.88	0.46	0.43	0.52	0.51	0.48	0.43
		both	0.43	**0.41**	0.87	0.45	0.43	0.51	0.51	0.48	0.42
5	0.20	none	**0.51**	**0.51**	1.06	0.54	0.53	0.63	0.62	0.58	0.53
		first	**0.48**	0.49	0.98	0.50	0.50	0.58	0.57	0.54	0.50
		last	**0.52**	0.54	1.13	0.56	0.56	0.66	0.66	0.62	0.55
		both	**0.49**	0.53	1.06	0.51	0.53	0.61	0.60	0.57	0.53
5	0.50	none	0.67	**0.66**	1.21	0.69	0.71	0.79	0.78	0.75	0.71
		first	0.56	**0.31**	0.42	0.45	0.37	0.42	0.42	0.40	0.35
		last	**0.66**	0.67	1.22	0.68	0.72	0.80	0.79	0.76	0.71
		both	0.59	**0.33**	0.44	0.46	0.39	0.44	0.44	0.43	0.37
6	0.05	none	0.79	**0.63**	1.39	0.77	0.69	0.71	0.71	0.70	0.71
		first	0.78	**0.60**	1.32	0.75	0.67	0.69	0.68	0.68	0.67
		last	0.80	**0.63**	1.37	0.78	0.69	0.72	0.71	0.70	0.71
		both	0.80	**0.61**	1.30	0.77	0.67	0.70	0.69	0.69	0.69
6	0.20	none	0.97	**0.64**	1.16	0.89	0.66	**0.64**	**0.64**	**0.64**	0.72
		first	0.94	**0.52**	0.95	0.81	0.55	0.53	0.53	0.53	0.59
		last	1.02	0.64	0.96	0.91	0.58	**0.56**	**0.56**	**0.56**	0.68
		both	0.97	0.51	0.80	0.81	0.49	**0.46**	**0.46**	**0.46**	0.55
6	0.50	none	1.28	0.61	0.85	1.02	0.53	**0.50**	**0.50**	0.51	0.59
		first	1.22	0.26	0.42	0.62	0.27	**0.25**	**0.25**	**0.25**	0.27
		last	1.39	0.62	0.69	1.05	0.45	**0.42**	**0.42**	0.43	0.53
		both	1.31	0.26	0.38	0.62	0.25	**0.22**	**0.22**	**0.22**	0.26

Table 5.9. Simulation study results of benchmark and Bounded Weights methods forecast combination methods for correlation matrices CM4, CM5 and CM6 (out-of-sample analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

Comparing BW(·) and LB, the latter is slightly more favorable with its smaller average rank and, looking at Figure 5.10, it has wider variety of ranks that is favorable as it is closer to rank one.

Bounded prior weights BW($\dot{\omega}$), in particular towards the shrinkage directions $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$, are the best method more often compared to BW and LB. Both $\boldsymbol{\omega}^{E3}$ and $\boldsymbol{\omega}^{IL}$ have similar ranks, 4.62 and 4.77 respectively, but they are larger than those of BW(·) and LB. $\boldsymbol{\omega}^{E2}$ has a noticeably larger rank with 5.41. Figure 5.10 reveals that although $\boldsymbol{\omega}^{IL}$ is more consistent (smaller IQR), $\boldsymbol{\omega}^{E3}$ can achieve smaller ranks. For the

	EW	PW	OW	IL	BW(·)	BW($\dot{\omega}$)			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
Smallest MSE (%)	23.61	**41.67**	0.00	29.17	4.17	22.22	23.61	15.28	4.17
Avg Rank	5.42	**3.35**	8.31	4.48	4.49	5.41	4.62	4.77	4.14
Avg Distance	0.27	0.07	0.50	0.18	**0.05**	0.07	0.06	0.06	**0.05**

Table 5.10. Key figures for the MSE values of benchmark and Bounded Weights methods over all simulation study scenarios (out-of-sample analysis). Smallest MSE (%) — Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank — Average rank of a method where a smaller rank is favorable. Avg Distance — Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

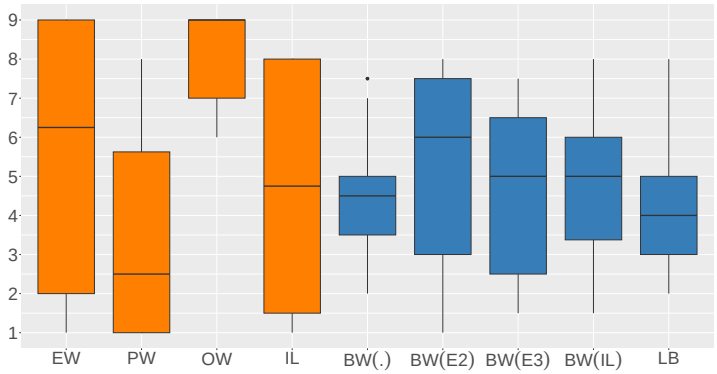


Figure 5.10. Boxplot of ranks across the benchmarks and Bounded Weights methods for the pseudo out-of-sample analysis.

former, in at least 25% of scenarios the rank is 3.5 or smaller while for the latter the corresponding value is about 2.5.

Comparing all BW(·), LB with the extended range of feasible values is preferable. It is relatively consistent while still achieving low ranks for some scenarios. This is strong evidence that the uncertainty introduced by hyperparameter estimation is a crucial factor. This is due to the fact that BW(·) nests LB and, thus, the former is at least as good as LB if one could estimate hyperparameters precisely and accurate.

If we consider all Bounded Weights methods as a whole they have a strictly smaller MSE for 16 scenarios (about 20% of scenarios) and an equal MSE for another 2 scenarios. Accordingly, our proposed method improve forecast accuracy out-of-sample. However, their potential is limited by the hyperparameter estimation. With no further information the PW approach is the most promising. It has the smallest MSE most often (about 30 scenarios) and has the smallest average rank and even though the distribution of ranks is larger, it is still in line with the other methods.

5.3.2.2 Out-Of-Sample: Groups of Scenarios

In this subsection we analyze the out-of-sample forecasting performance of the forecast combination methods with respect to the correlation matrices, error variance similarities and special groups.

Table 5.11 shows how often each method has the smallest MSE (potentially among other methods) for each correlation matrix

	EW	PW	OW	IL	BW($\cdot$)	BW($\dot{\omega}$)			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
CM1	0.00	33.33	0.00	0.00	25.00	**58.33**	**58.33**	33.33	25.00
CM2	33.33	33.33	0.00	**50.00**	0.00	16.67	25.00	16.67	0.00
CM3	41.67	16.67	0.00	**75.00**	0.00	0.00	0.00	0.00	0.00
CM4	25.00	**50.00**	0.00	**50.00**	0.00	0.00	0.00	0.00	0.00
CM5	41.67	**66.67**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CM6	0.00	50.00	0.00	0.00	0.00	**58.33**	**58.33**	41.67	0.00

Table 5.11. Percentage of scenarios for which benchmarks and Bounded Weights methods have the smallest MSE with respect to the error correlation matrix (out-of-sample analysis). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

BW($\dot{\omega}$) with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ have the best out-of-sample MSE values for highly correlated forecast errors (CM1 and CM6) most often. This is similar to what we observed for the ex post analysis in Table 5.6. For the ex post analysis, BW($\cdot$) was always superior for CM4 and CM5, i.e., with mixed error correlations. However, for pseudo out-of-sample forecasting the only three scenarios for which BW, as well as LB, have the smallest MSE are for CM1, i.e., highly correlated forecast errors. Figure 5.11 again presents the average ranks and distances to the best method. BW($\dot{\omega}$) has smaller ranks and distances than BW($\cdot$) and LB for both CM1 and CM6.

With respect to the benchmark methods EW is never the superior method for CM1 and CM6. Based on Figure 5.11 we can see that for CM1 BW($\dot{\omega}$) with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ and for CM6 BW($\dot{\omega}$) in general have the smallest rank, i.e., most consistently small MSE. This is somehow surprising as PW is oftentimes the superior method for both CM1 (33.33%) and CM6 (50%). However, both its average rank and distance is inferior compared to the BW($\dot{\omega}$) methods. If one has only information that forecasts are highly correlated, the BW($\dot{\omega}$) methods should be used. With respect to BW($\cdot$) and LB it reveals that over all correlation matrices beside CM1 and CM6, LB has a smaller average rank.

From high forecast error correlations (0.9, CM1), to medium (0.5, CM2) and low correlations (0.2, CM3) the inverse-loss weighted average is more suited as it has. While for CM2 PW is preferable with regard to its average rank, IL has the smallest MSE

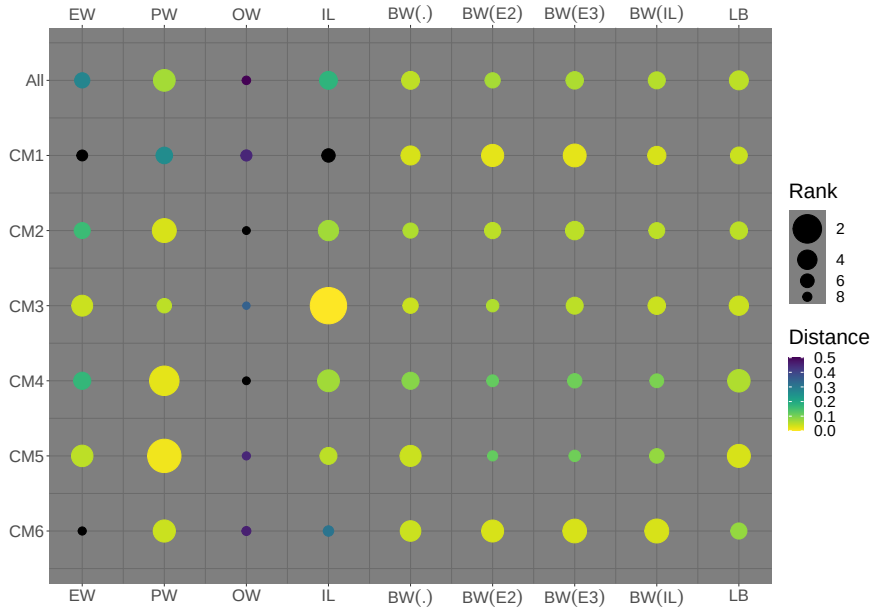


Figure 5.11. Illustration of average ranks and distances of the benchmarks and Bounded Weights for different correlations matrices (out-of-sample analysis).

more often. For CM3 IL is the best method most often, has the smallest average rank and distance, see Table 5.11 and Figure 5.11. While PW is more often the best method for CM3, BW($\dot{\omega}$) with $\boldsymbol{\omega}^{E3}$ and $\boldsymbol{\omega}^{IL}$ as well as LB have on average smaller ranks, i.e., a better forecasting performance.

Recall, that for CM4 forecast are homogeneous within groups and for CM5 the better the forecast accuracy the higher the amount of error correlation, see again Section 3.2.2.2. Only with respect to all Bounded Weights methods (BW($\cdot$)), BW($\cdot$) and LB are better suited for CM4 and CM5 as they have smaller ranks on average. However, if we also consider the PW approach, it has a better forecasting performance overall. Moreover, it has the smallest MSE more often and a smaller average rank and distance.

Table 5.12 presents how often each method has the smallest MSE with respect to the error variance similarities and special groups. In addition to that Figure 5.12 depicts the average ranks and distances. For the error variance similarities one can see that the BW($\dot{\omega}$) methods are better more often, the less similar the error variances of forecasts are. In particular the shrinkage direction we proposed to use $\boldsymbol{\omega}^{E3}$ has the smallest MSE most often out of all Bounded Weight methods. Similarly, the average ranks of BW($\dot{\omega}$) get better, the more dissimilar the forecast error variances are, see Figure 5.12. The contrary holds for the BW($\cdot$) and LB. According to the average ranks they are more suited for similar forecast error variances.

EW, IL and in particular PW are, however, favorable for similar forecast error variances as they have the smallest MSE more often, better average ranks and distances, see Table 5.12 and Figure 5.12. With increases error variance similarity a decreasing

	EW	PW	OW	IL	BW(·)	BW			LB
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
$z = 0.05$	45.83	50.00	0.00	**45.83**	0.00	0.00	0.00	0.00	0.00
$z = 0.20$	20.83	29.17	0.00	**33.33**	8.33	29.17	29.17	20.83	8.33
$z = 0.50$	4.17	45.83	0.00	8.33	4.17	37.50	**41.67**	25.00	4.17
none	27.78	**44.44**	0.00	33.33	0.00	22.22	16.67	5.56	0.00
first	22.22	**50.00**	0.00	22.22	5.56	11.11	22.22	22.22	5.56
last	27.78	27.78	0.00	**38.89**	5.56	27.78	27.78	5.56	5.56
both	16.67	**44.44**	0.00	22.22	5.56	27.78	27.78	27.78	5.56

Table 5.12. Percentage of scenarios for which the benchmarks and Bounded Weights methods have the smallest MSE with respect to the error variance similarity and special groups (out-of-sample analysis). The total number of scenarios for each error variance similarity is 24 and for special groups it is 18. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

forecasting performance of EW is expected and can be observed. With the introduction of larger error variance, see again Figure 3.5, the performance of an approach that uses all forecasts and averages them has to decline. Overall, the average ranks of EW and IL decrease as the forecast error variances become more dissimilar ($z = 0.2$ and $z = 0.5$). For PW the ranks are better for $z = 0.5$ compared to $z = 0.2$ but for both it has better average ranks than for $z = 0.05$. Although, IL has a similar rank to PW for 0.2 and is the best method more often, it has a noticeably larger average distance to the best method.

In conclusion, overall for all degrees of forecast error variances similarity, the PW approach is favorable due to the proportion how often it is the best method and, in particular, with regard to the average rank.

For the special groups, again, there is less structure in the percentages of being best, average rank and distance. For SG none, first and both PW is the best method most often and has the best ranks. However, the average distance is in fact smaller for BW(·): Although PW is oftentimes the best method, if it is not, the deviation in MSE from the corresponding best method is larger. This hold similar for IL. It has similar average ranks than BW(·), but its average distance is noticeably larger, see Figure 5.12.

5.3.3 Summary of Results

Overall the Bounded Weights methods and in particular BW(·) have a noticeably superior forecasting performance within the ex post analysis in Section 5.3.1. While BW($\dot{\omega}$) are favorable if forecasts have a high error correlation, the BW(·) approach is more suited for smaller and mixed error correlations. If forecast errors have similar variances, BW(·) is always the best choice. The more diverse forecasts are in terms of their error variance, the better BW($\dot{\omega}$). Nevertheless, BW(·) is still a competitive fore-

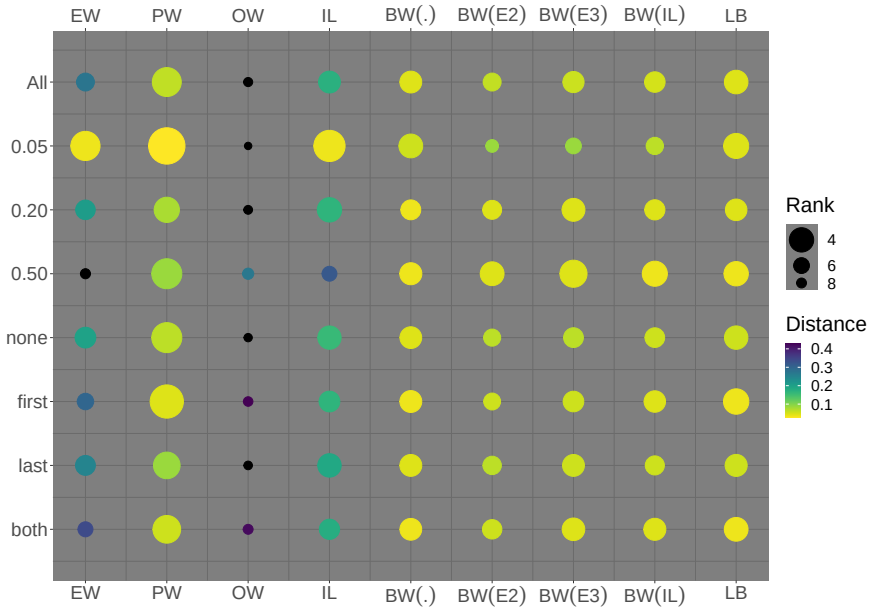


Figure 5.12. Illustration of average ranks and distances of the benchmarks and Bounded Weights methods for error variance similarities and special groups (out-of-sample analysis).

cast combination method for these scenarios. With respect to special groups there is no noticeable difference in forecast performance for any method. For all 72 scenarios there is a Bounded Weights methods that either has a smaller MSE than all benchmark methods or the same MSE. For 67 scenarios the MSE of at least one Bounded Weights methods has a strictly smaller MSE than the benchmarks.

If hyperparameter have to be estimated, the clear superiority of Bounded Weights methods is not given. For out-of-sample forecasting, if no further information is available the PW approach is the most favorable, because it has the smallest MSE most often, average rank and distance in general. In contrast to the ex post analysis Bounded prior weights BW($\dot{\omega}$) are in general favorable over BW($\cdot$) and LB. This holds in particular, if we use Conditional Group Equal Weights, i.e., the shrinkage directions we suggest to use within the optimization framework $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$.

With respect to the different error correlation matrices, the results suggest that for highly correlated forecast errors (CM1 and CM6) BW($\dot{\omega}$) is favorable. For medium correlations, homogeneous forecasts within groups and if the amount of correlation depends on the forecast accuracy (CM2, CM4 and CM5) PW is the best choice within this analysis and hyperparameter estimation procedure. For small error correlations (CM3) IL has the best performance overall. If we consider the different forecast error variances, the BW($\dot{\omega}$) methods are favorable if the forecast error variance are more dissimilar while the opposite is true for BW($\cdot$) and LB. The PW is most favorable for similar forecast error variances but also the preferable method as the forecast error variances become

more dissimilar. For special groups PW is most often the best method and, if not, close to the best method in terms of its rank. However, BW(·) is oftentimes closer with respect to the difference in MSE. Accordingly, sometimes PW has a noticeably larger MSE than the best method, although it usually ranks close to it. If we consider all Bounded Weights methods as a whole they have a strictly smaller MSE for about 20% of scenarios and an equal MSE for another 2 scenarios.

5.4 Discussion and Future Work

The objective of this chapter was to introduce new methods for forecast combination with constrained weights. These methods impose constraints on the weights in order to shrink them and, by that, improve the out-of-sample forecasting performance. As demonstrated in Chapter 4 forecast combination with L_1 constraints shrinks weights towards fixed values or prior weights. To this end, a single constraint is imposed that restricts the weight vector as a whole. This approach does not explicitly prevent large (small) weights from having a large effect or impact on the combined forecast, relative to other weights. In other words, the combined forecast more strongly depends on a few forecasts and, therefore, it is less robust and diversified.

With respect to the overall structure of this thesis, this chapter present major contribution (IV) stated in Chapter 1: Forecast Combination with Bounded (Prior) Weights. We present a new forecast combination method with constrained weights that improves the forecast accuracy compared to the benchmarks. This directly contributes to our overarching research question: *how to further improve the forecast accuracy of a combined forecast using constrained weights?*

(I) In Section 5.1.1 we proposed a new method: Forecast Combination with Bounded Weights (BW). To this end, we utilize lower and upper bounds to shrink weights and restrict the overall size of them, i.e., the effect one forecast has on the combined forecast. By introducing both common lower and upper bounds into the forecast combination method of J. M. Bates and Granger (1969) we developed a method that nests the PW, OW and EW approach as well as solutions in-between them. A related approach from Radchenko et al. (2023) used only a lower bound to limit the amount of negativity. However, in the way it was implemented it only nests PW and OW but not EW. By re-evaluating the feasible bounds for BW, we extended the interval of feasible values for the lower bound to positive values smaller or equal to EW. This extension allows for the incorporation of the EW solution even if only a lower bound is used.

(II) In Section 5.1.2 we showed analyzed solutions of BW(·). The solutions in-between the nested methods EW, PW and OW can have groups of weights: individual and identical weights. If either the lower or upper bound is set to equal weights, all weights are identical. As the lower (upper) bound decreases (increases), some weights deviate

from the identical solution while others are equal to the current lower (upper) bound. Depending on the values of the bounds, there can be two groups of identically weighted forecasts. The identical weights of the groups are then equal to the lower or upper bound respectively. Accordingly, our approach allows for some weights to have individual weights while others have identical weights. In this way, our approach is related to the EW approach. It creates a diverse solution such that the combined forecast more robust. Simultaneously, we capture potential for improvement by letting some weights deviate from the identical weights.

In Section 5.2 we extended to the BW($\cdot$) approach to incorporate prior information in the form of weights (objective III of this chapter): Forecast Combination with Bounded Prior Weights (BW($\dot{\omega}$)). We defined individual bounds for each weight, which were determined by a common deviation from the prior weights. To this end, we used the same prior weights or shrinkage directions as in Chapter 4, i.e., $\boldsymbol{\omega}^{E2}$, $\boldsymbol{\omega}^{E3}$ (Conditional Group Equal Weights) and $\boldsymbol{\omega}^{IL}$. For example, recall that for $\boldsymbol{\omega}^{E2}$ we have two groups of forecasts with equal weights conditional to their assigned budget. By defining bounds based on a common lower and upper deviation from the prior weights, solutions have similar behavior to BW($\cdot$), but for two groups. In other words, the bounds shrink weights to the conditional equal weight of their corresponding group. Each group can have identical and individual weights.

For both BW($\cdot$) and BW($\dot{\omega}$) there are two hyperparameters that must be estimated, e.g., by cross-validation with a grid search. It is important to note, that the possible values of one bound or deviation depend on the value of the other bound or deviation respectively. To ensure an efficient hyperparameter estimation, we introduced an algorithm that evaluates only the necessary pairs of potential lower and upper bounds or deviations in Section 5.1.3.

(IV) In Section 5.3 we evaluated the forecast performance of BW($\cdot$), BW($\dot{\omega}$) and LB in terms of an ex post and out-of-sample analysis. We summarized the results in Section 5.3.3. In the ex post analysis the Bounded Weights methods and in particular BW($\cdot$) were by far superior to other forecast combination methods. Shrinkage towards prior weights BW($\dot{\omega}$) is more suited for highly correlated or forecasts with diverse error variances, i.e., forecasting performances, whereas BW($\cdot$) is better suited for smaller and mixed error correlations as well as more similar forecast error variances.

For 67 out of 72 scenarios of the simulation study the Bounded Weights methods (BW($\cdot$) and BW($\dot{\omega}$)) had a strictly smaller MSE than the benchmarks. For each of the remaining five scenarios one of the benchmark methods had the same MSE as one of the Bounded Weights methods.

For about 20% of scenarios in the out-of-sample analysis one of the Bounded Weights methods has a MSE strictly smaller than the benchmarks. However, for many scenarios, the PW approach was a strong benchmark. With respect to the LB method, there is

evidence that the feasible values for the lower bound should include to positive values. Then, the method is better suited for scenarios in which a solution closer to equal weights is preferable.

The uncertainty introduced by the hyperparameter estimation leads to a lower forecasting performance of the Bounded Weights methods. In theory, compared to the ex post analysis the performance of methods that have hyperparameters can improve if we re-estimate the hyperparameters for each observation in the test set as we did for the out-of-sample analysis. However, the observed results indicate that the hyperparameter estimation is not accurate enough. This is amplified by the fact that we use relatively short time series, i.e., there is only a small amount of data. Additionally, the number of weights (24) that have to be estimated is considerable in comparison. Both aspects increase estimation uncertainty. Similar to the findings from Chapter 4 for the L_1 method we see the necessity for future research that focuses on the hyperparameter estimation. As a part of that, the analysis can be repeated for a larger number of observations. Additionally, it would be beneficial to analyze which bounds where used in the different scenarios. Based on these results one could perform a more informed search for the hyperparameters or identify appropriate bounds that can be used for a range of scenarios.

In conclusion, the utilization of prior weights BW($\dot{\omega}$) allows for the incorporation of prior information in the form of shrinkage directions. We impose bounds around them that allow for some deviations of single weights, while retaining the idea or direction of the prior weights. For instance, two groups of forecasts with different conditional equal weights ($\boldsymbol{\omega}^{E2}$). Our proposed approach of using both a lower and upper bound BW($\cdot$) allows for solutions in-between (EW, PW and OW). This approach combines the advantages of the three methods while mitigating their flaws. By imposing constraints on all weights individually, our solutions has groups of individual and identical weights which leads to a more diversified and, by that, robust combination of forecast that is inspired by the historically tough benchmark: equal weights.

6 Individual Feature Bounds

The Bounded Weight method from the previous chapter directly contributed to the overarching research question of this thesis: *how to further improve the forecast accuracy of a combined forecast using constrained weights?* However, we stated another research question in Chapter 1: *how to incorporate additional, external information in forecast combination with constrained weights?* In this chapter, we extend the bounded weights approach by incorporating additional information to constrain weights and, thus, consider both research question of this thesis.

In theory the best possible weights to combine forecast are the weights of the original forecast combination problem by J. M. Bates and Granger (1969) depicted in Equation (2.22). However, due to the forecast combination puzzle, see again Section 2.3, other methods have been developed to determine weights and thereby combine forecasts. This includes variants that are based on the OW approach. The PW approach uses only positive weights, see Equation (2.30). In Chapter 4 we introduced several methods that use shrinkage to constrain weights in order to improve the out-of-sample forecast accuracy. This includes the linear (hybrid) shrinkage and L_1 constrained method. The former shrinks weights from the OW solution towards the equal weights solution. In case of LHS, a subset is shrunken to zero. L_1 constraint shrink weights towards zero, equal weights or other prior weight vectors like the inverse-loss weighted average or the shrinkage directions we proposed to use, i.e., groups of weights are shrunken towards their conditional equal weight defined by an assigned budget. In Chapter 5 we introduced forecast combination with bounded weights. It imposes bound constraints that ensure that the influence or marginal effect of a forecast on the combined forecast is restricted and, thereby, the solution is more diversified while simultaneously allowing for individual weights to improve the forecast accuracy. Based on this, we presented a method that imposes bounds around any prior weights vector, i.e., we shrink towards these weights. All approaches mentioned are based on the forecast combination problem defined by J. M. Bates and Granger (1969). The objective function is to minimize the in-sample error variance or MSE under constraints, at least the unity constraint. Constraints are imposed based on parameters like γ_k in case of the L_1 constraint or $\underline{\omega}$ and $\overline{\omega}$ in case of the Bounded Weights methods. However, there are other ways to determine weights.

The most prominent example is the equal weights forecast. Another already considered way to determine weights is by the inverse-loss weighted average of Equation (4.16).

Alternatively, weights can be determined by other features than the MSE. For example, Kolassa (2011) combine forecasts weights based on information criterion like the Akaike information criterion. Specific traits or features as the accuracy and diversity of forecasts can be used to determine weights and combine forecasts (Davis-Stober, Budescu, Broomell, & Dana, 2015; Merkle, Saw, & Davis-Stober, 2020). Montero-Manso et al. (2020) proposed an automated method FFORMA (Feature-based Forecast Model Averaging) that estimates weights based on 42 characteristics of the time series itself. It uses a meta-learning approach in form of a gradient tree boosting model from xgboost (Chen & Guestrin, 2016). A meta-learning approach learns across multiple time series instead of one. FFORMA was the second most accurate contribution to the M4 Competition (Makridakis et al., 2018, 2020) both for point forecasts and prediction intervals (Montero-Manso et al., 2020; Wang et al., 2023). There are even more combination schemes, for example for interval forecasts (Wang, Kang, & Li, 2022), within Bayesian forecast combination (Li, Kang, Petropoulos, & Li, 2023) or for intermittent demand (Li, Kang, & Li, 2023).

We took inspiration from the forecast combination problem with L_1 constraint, the Bounded Weights methods and the feature-based combination schemes and introduce a new field within forecast combination. While weights so far have been either constrained based on fixed values or determined based on different features, we propose to constrain weights based on feature values. In general, constraints improve forecast accuracy, see again Chapters 4 and 5. However, the question is whether or not a constraint concerns either all forecast together (L_1) or individually but with the same bounds or deviation (BW). In other words, the main idea is that forecasts that have a favorable feature value are less constraint than forecast with less favorable feature values. Thereby, we utilize the positive effects of the considered shrinkage methods, i.e., more similar marginal effects and improved forecast accuracy. Importantly, while allowing some favorable forecasts to be less constraint to improve the overall out-of-sample forecast accuracy. Constraints that are determined by feature values of the forecasts introduce exogenous, additional information into the forecast combination problem. This exogenous information allows us to adapt our method for different applications. For example, if we consider forecasting in the possible presence of economic shocks, we can define a feature that measures how fast a forecast adapts to short-term changes such that the combined forecast more quickly captures changes in the time series. Assume we want to combine forecasts for food retail but are in particular interested in a good forecast performance during promotional periods. We can evaluate each input forecast only for promotional periods and incorporate the results as a feature based on which we derive the individual feature bounds. Thereby, we constrain forecasts less that are more suitable for promotional periods.

Within this chapter we introduce *Forecast Combination with Individual Feature Bounds*. At its core it is an extension of the Bounded Weights methods. However, it is a new field of research that applies the concept of bounds but allows for exogenous information to be used to define the constraints for specific applications, situation, and structures of the data and, ultimately, improve forecast accuracy.

The objective of this chapter are the following:

(I) Introduce our new approach of Forecast Combination with Individual Feature Bounds

(II) Demonstration of the versatility of this methods by presenting various features that can be used to define bounds.

(III) Analyze the forecasting performance of our new method within our simulation study both in an ex-post and out-of-sample analysis.

With respect to the overall structure of this thesis, the introduction of Forecast Combination with Individual Feature Bounds is the fifth main contribution we stated in Chapter 1. It not only contributes to the overarching research question of how to further improve forecast accuracy of forecast combination with constrained weights. Additionally, it provides a new method that allows to incorporate additional, external information which was the second research question of this thesis.

The remainder of this chapter is organized as follows. Section 6.1 introduces the concept of individual feature bounds, provides a formal definition of the optimization problem and discusses its core components. Thereafter, Section 6.1.2 presents different *transformation functions* that are used to map the feature values to the individual constraints. Section 6.1.3 presents different features of forecasts that can be used. Lastly, in Section 6.1.4 we discuss our results and directions for future research.

6.1 Individual Feature Bounds

In the previous chapters, we used prior weights for both the L_1 methods and Bounded Weights methods, see Chapters 4 and 5. Prior weights allow for a more general representation of the methods. For example, we can define all prior weights to be zero and, as a result, if we consider the L_1 constraint we have $L_1(0)$ while for BW($\boldsymbol{\omega}$) we get BW. Therefore, for forecast combination with individual feature bounds we use prior weights to allow for the most flexible forecast combination method. For example, we can use the conditional group equal weights shrinkage directions $\boldsymbol{\omega}^{E2}$ or $\boldsymbol{\omega}^{E3}$ shown in Section 4.2.3.

The forecast combination problem with individual feature bounds (IFB) is to an extent similar to BW($\boldsymbol{\omega}$) of Equation (5.9). For BW($\boldsymbol{\omega}$) we used common deviations

from prior weights, $\underline{\beth}$ and $\overline{\beth}$, to determine the individual lower and upper bounds, $\underline{\omega}_i$ and $\overline{\omega}_i$, see again Equations (5.10) and (5.11). We used common deviations because individual lower and upper bounds for each prior weight are at least impractical or even impossible if all $2N$ bounds have to be estimated by cross-validation, see again Section 5.2.

For forecast combination with IFB, we use a similar looking formula as for BW($\dot{\boldsymbol{\omega}}$) to determine the individual feature lower and upper bounds:

$$\underline{\omega}_i = \dot{\omega} - \underline{\aleph}_i \qquad \forall\, i = 1, \ldots, N, \tag{6.1}$$

$$\overline{\omega}_i = \dot{\omega} + \overline{\aleph}_i \qquad \forall\, i = 1, \ldots, N. \tag{6.2}$$

$\aleph_i$ is the *Individual Feature* ***Deviation*** *(IFD)* from prior weights $\dot{\omega}_i$.[43] The lower IFDs are denoted by $\underline{\aleph}_i$ and the upper IFDs by $\underline{\aleph}_i$. Henceforth, the bounds in Equations (6.1) and (6.2) are the method defining *Individual Feature* ***Bounds*** *(IFB)*. In order to determine them, $\underline{\aleph}_i$ and $\overline{\aleph}_i$ are used as a deviation from prior weights and are denoted by IFD.

In contrast to the common feature deviations ($\underline{\beth}_i, \overline{\beth}_i$) from BW($\dot{\boldsymbol{\omega}}$), $\underline{\aleph}_i$ and $\overline{\aleph}_i$ (IFD) are individual for each prior weight or forecast i. The IFDs are determined based on features or characteristics of the forecasts $\hat{y}_i$ $\forall i = 1, \ldots, N$. We will discuss the IFDs in more detail later, but for now we assume that $\underline{\aleph}_i$ and $\overline{\aleph}_i$ are given for each weight or forecast.

Based on Equations (6.1) and (6.2), we can replace both $\underline{\omega}_i$ and $\overline{\omega}_i$ from the forecast combination problem with bounded prior weights BW($\dot{\boldsymbol{\omega}}$) depicted in Equation (5.9) by the right-hand side of Equations (6.1) and (6.2). This leads to the forecast combination problem with individual feature bounds (IFB):

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \omega_i - \dot{\omega}_i \geq -\underline{\aleph}_i \qquad \forall\, i = 1, \ldots, N, \\ & \omega_i - \dot{\omega}_i \leq \overline{\aleph}_i \qquad \forall\, i = 1, \ldots, N \end{aligned} \tag{6.3}$$

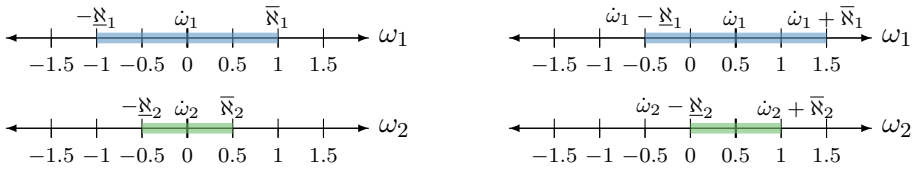
with $\underline{\aleph}_i, \overline{\aleph}_i \in \mathbb{R}_{\geq 0}$. For each forecast $\hat{y}_i$, we constrain its weight to be within an interval around a prior weights $\dot{\omega}_i$, i.e.,

$$\dot{\omega} - \underline{\aleph}_i \leq \omega_i \leq \dot{\omega} + \overline{\aleph}_i \qquad \forall\, i = 1, \ldots, N. \tag{6.4}$$

In other words, we introduce both a lower and upper bound for each weight that depends on an individual deviation from its prior weight defined by $\underline{\aleph}_i$ and $\overline{\aleph}_i$.

[43] Alef $\aleph$ is a letter in the Hebrew alphabet.

For example, let there be two forecasts, i.e., $\hat{y}_i \; \forall i \in \{1, 2\}$. The IFDs are $\underline{\aleph}_1, \overline{\aleph}_1 = 1$ and $\underline{\aleph}_2, \overline{\aleph}_2 = 0.5$. Figures 6.1(a) and 6.1(b) illustrate the interval of feasible values of weights for ω_1 on top and ω_2 on the bottom for two different prior weights vectors.



(a) $\dot{\omega}_1 = \dot{\omega}_2 = 0$ **(b)** $\dot{\omega}_1 = \dot{\omega}_2 = 0.5$

Figure 6.1. Illustration of individual feature bounds around $\dot{\omega}_i$ defined by the individual feature deviations $\underline{\aleph}_i$ and $\overline{\aleph}_i$. Note that $\dot{\omega}_2$ in Figure 6.1(b) is not depicted for the sake of better readability.

In Figure 6.1(a) the prior weights are $\dot{\omega}_1 = \dot{\omega}_2 = 0$, i.e., this corresponds to shrinkage towards zero. Figure 6.1(a) depicts the interval of feasible values for ω_1, i.e., $[-\underline{\aleph}_1, \aleph_1]$ by the blue rectangle. ω_1 has to be between -1 and 1. Because both individual feature deviations $\underline{\aleph}_1, \overline{\aleph}_1$ are identical, ω_1 can have values symmetrically around zero. For ω_2, the individual feature deviations are smaller with 0.5, i.e., the feasible interval depicted by the green rectangle is tighter around zero. Thereby, ω_2 is forced to be closer to zero. For $\dot{\omega}_1 = \dot{\omega}_2 = 0.5$ (equal weights) in Figure 6.1(a) both intervals of feasible values shift to the right, ceteris paribus. They are centered around equal weights, i.e., 0.5. While ω_1 can have a larger deviation from equal weights, ω_2 has to be closer to it. Keep ind mind that later we will determine the amount of deviation ($\underline{\aleph}_i$ and $\overline{\aleph}_i$) based on feature values of the forecasts.

If individual prior weights are used for both forecasts, the bounds are centered around these prior weights. For example, let $\dot{\omega}_1 = 0$ and $\dot{\omega}_2 = 0.5$ the corresponding intervals for ω_1 is given by the blue rectangle of Figure 6.1(a) and for ω_2 it would be the green rectangle of Figure 6.1(b).

If the individual feature deviations $\underline{\aleph}_i, \overline{\aleph}_i$ are not identical, an exemplary interval of feasible values is depicted by the following figure.

Figure 6.2. Illustration of asymmetrical individualized feature bounds around $\dot{\omega}_1 = \dot{\omega}_2 = 0$. Note that $\dot{\omega}_2$ is not depicted for a sake of better readability.

In this case, the interval of feasible values is asymmetrical around $\dot{\omega}_1 = \dot{\omega}_2 = 0$.

Lastly, one can only use either the lower IFB or upper IFB. By that, one defines a half-closed interval of feasible values for each weight, i.e., $\omega_i \in [\dot{\omega}_i - \underline{\aleph}_i, \infty)$ or $(-\infty, \dot{\omega}_i + \underline{\aleph}_i)$

respectively. The following figure depicts the feasible interval if only the lower IFB is used.

Figure 6.3. Illustration of the half closed feasible intervals for two weights if only lower IFBs, i.e., IFDs $\underline{\aleph}_i$, are used with $\dot{\omega}_i = 0 \; \forall \; i = 1, \ldots, N$.

6.1.1 Idea and Components of Individual Feature Bounds

To introduce forecast combination with individual feature bounds we previously assumed that the individual feature deviations (IFDs) $\underline{\aleph}_i$ and $\overline{\aleph}_i$ that define the individual lower and upper bound (IFBs) $\underline{\omega}$ and $\overline{\omega}$ are given. In what follows, we introduce the framework how $\underline{\aleph}_i$ and $\overline{\aleph}_i$ are determined.

We defined both $\underline{\aleph}_i$ and $\overline{\aleph}_i$ to be non-negative, see again Equation (6.3). Although, $\underline{\aleph}_i$ and $\overline{\aleph}_i$ are the lower and upper individual feature deviation, they are defined in the same way. Therefore, for the sake of simplicity we only use $\aleph_i$, however, the following concepts are valid for both $\underline{\aleph}_i$ and $\overline{\aleph}_i$.

Let $\aleph_i$ be the i-th element of the $N \times 1$ of an *individual feature deviation vector*, $\boldsymbol{\aleph} = (\aleph_1, \ldots, \aleph_N)'$. The IFD vector $\boldsymbol{\aleph}$ is determined by

$$\boldsymbol{\aleph} = \Psi\Big(\boldsymbol{\nu}, \psi_{min}, \psi_{max}, ...\Big). \tag{6.5}$$

We call Ψ a *transformation function* as it transforms or maps an input to the IFD $\boldsymbol{\aleph}$. The input of the transformation function is, inter alia, the $N \times 1$ *feature vector* $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)'$ with

$$\nu_i = \xi(\hat{\boldsymbol{y}}_i) \qquad \forall \; i = 1, \ldots, N. \tag{6.6}$$

The $N \times 1$ vector $\hat{\boldsymbol{y}}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,\tau})'$ includes forecasts i for all time periods within the training set, i.e., $t = 1, \ldots, \tau$. The term ξ is a function that calculates the feature values based on the in-sample data.

Before we take a closer look at the components of the transformation function of Equation (6.5), we want to summarize the idea of our approach again. There is feature vector $\boldsymbol{\nu}$ that is determined based on the forecasts, and it reflects a feature, characteristic or key performance indicator of them. To put it differently, it reflects how each forecast is evaluated in context of a certain criterion. Each feature value ν_i for every forecast is transformed or mapped by the transformation function Ψ to an IFD $\aleph_i$ which is then used together with the prior weights $\dot{\omega}_i$ to determine the lower or upper IFB for weight $\omega_i \; \forall \; i = 1, \ldots, N$ in Equation (6.3).

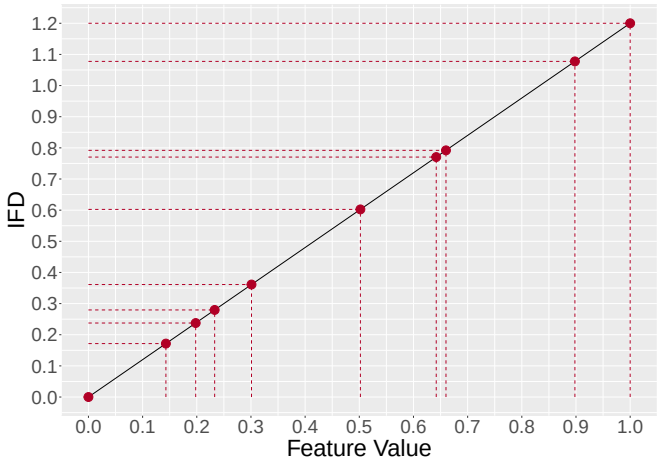


Figure 6.4. Illustration of the relationship between the feature vector $\boldsymbol{\nu}$, a linear transformation function Ψ and the Individual Feature Deviation $\boldsymbol{\aleph}$. The values for $\boldsymbol{\nu}$ where randomly chosen between zero.

Figure 6.4 depicts how the transformation functions Ψ maps the feature vector $\boldsymbol{\nu}$ to the IFD vector $\boldsymbol{\aleph}$. We use a simple linear transformation function Ψ depicted by the black line exemplary. The feature values ν_i of an arbitrary feature vector $\boldsymbol{\nu}$ are on the abscissa and the IFDs $\aleph_i$ are on the ordinate. Each value of $\boldsymbol{\nu}$ is represented as a red dot on the transformation function, i.e., the black line. The red, dashed lines function as visual aid to connect the feature values ν_i and the corresponding IFD $\aleph_i$. For any feature value ν_i one can determine $\aleph_i$ using Figure 6.4. For example, the forecast $(\hat{y}_j)$ with a feature value of roughly 0.5 has an IFD of 0.6, i.e., it can deviate from its prior weights by that amount. If we assume that $\underline{\aleph}_j = \overline{\aleph}_j$ the corresponding weight ω_j with $j \in \{1, \ldots, N\}$ has the constraints

$$\omega_j \leq \dot{\omega}_j + 0.6, \tag{6.7}$$

$$\omega_j \geq \dot{\omega}_j - 0.6. \tag{6.8}$$

Note that within this example one could also simply scale the feature vector differently and, by that, impose $\boldsymbol{\nu} = \boldsymbol{\aleph}$. However, we want to provide a general framework for forecast combination with individual feature bounds that can be used for various transformation functions Ψ. We will discuss some transformation functions in Section 6.1.2.

The transformation function $\Psi\big(\boldsymbol{\nu}, \psi_{min}, \psi_{max}, ...\big)$ of Equation (6.5) has three specified arguments or components that all transformation function that we will use have in common. Depending on the choice of Ψ additional arguments are needed to define its shape. Those additional arguments are indicated by the three dots in Equation (6.5).

Feature values The first common argument of the transformation function is the feature vector $\boldsymbol{\nu}$, i.e., all feature values ν_i for each forecast. In Equation (6.6) it is defined by a function, ξ, which is a placeholder for the feature or criteria that is used. In the field of forecasting an obvious example for ξ is the forecast accuracy that can be measured by the MSE, recall Equation (2.2). In Equation (6.6) we use the in-sample forecasts as the input of ξ. However, one can easily adapt this and use different functions ξ that are based on other inputs than the forecasts. The function ξ can also be used to combine multiple features to one as we will discuss in more detail in Section 6.1.3.

With regard to the transformation function Ψ we have to account for the scale of its input, i.e., the scale of the feature vector. If the range or order of magnitude of the feature values changes, they will be mapped differently to $\aleph$. The scale of the feature values is first and foremost driven by the choice ξ. Accordingly, for different ξ, we would need to adapt the transformation functions to ensure that it has the desired properties for the inputs scale. However, even for the same ξ the feature values will in general vary, at least slightly, for two similar training sets during out-of-sample forecasting and cross-validation. If we want to impose properties for the transformation function, e.g., define the smallest and largest IFD in $\aleph$, we need to scale the feature vector $\boldsymbol{\nu}$ onto predefined range. In other words, we control the input values in Ψ to be within that predefined range and, thus, we can use the same transformation functions Ψ for different ξ and different training sets without the need to adapt it. To this end, we scale or normalize the feature vector $\boldsymbol{\nu}$ such that the most favorable feature value with is always equal to one and the least favorable feature value is always zero. Thereby, for all scaled feature values it holds that $\tilde{\nu}_i \in [0, 1]$.

By scaling the features, we also eliminate the problem that for some features the most favorable value might be the largest value, while for others smaller feature value are preferable. In that case we would need to adapt the transformation function accordingly to ensure that forecasts with more favorable features values are less constraint. However, when we scale the feature values, we can define the *scaled feature vector* $\tilde{\boldsymbol{\nu}}$ such that a larger value always indicates are more favorable feature value. To this end, if a larger feature value is more favorable we use

$$\tilde{\boldsymbol{\nu}} = \frac{\boldsymbol{\nu} - min(\boldsymbol{\nu})}{max(\boldsymbol{\nu}) - min(\boldsymbol{\nu})}. \tag{6.9}$$

If instead the smallest feature value is most favorable we use

$$\tilde{\boldsymbol{\nu}} = \frac{max(\boldsymbol{\nu}) - \boldsymbol{\nu}}{max(\boldsymbol{\nu}) - min(\boldsymbol{\nu})}, \tag{6.10}$$

(see e.g., Aggarwal, 2023, p.64). As a consequence, the forecast with least (most) favorable feature value will always have a feature value of zero (one).

As a consequence of scaling the feature vector, we can not compare the IFD vectors $\aleph$ from different sets of forecasts. Assume that we use the same feature determined by ξ and an identical transformation function Ψ for two sets of forecasts. Let one of the sets of forecasts have twice as large feature values than the other. Nevertheless, for both sets of forecasts the scaled feature vector $\tilde{\boldsymbol{\nu}}$ is zero for the least favorable and one for the most favorable feature value respectively. Although, forecast from one set have superior feature values, they will be constraint similarly to the inferior forecasts. For our purpose this is not a problem. We develop forecast combination with IFBs to combine forecast from a given set. We have no intention to compare the IFDs between sets of forecasts to gain insights about those sets.

We want to emphasize another important aspect that comes with scaling the feature vector: the unknown distribution of features values within it. The smallest and largest feature values of $\boldsymbol{\nu}$ are scaled to zero and one respectively. All other feature values are somewhere in-between zero and one. They can be more evenly distributed or heavily concentrated somewhere, including close to zero or one. By that, depending on the transformation function a larger proportion of forecasts can get similar IFDs which can lead to infeasible solutions, as we will discuss and solve later. As a consequence, it is more difficult to control the IFD by the design of the transformation function.[44] In other words, one can not guarantee that properties that follow from the design of Ψ are passed on to the IFD. Therefore, we suggest that ranks of the feature values can be used which solves this aspect. The most favorable feature value gets rank one and the least favorable feature value gets the largest rank. Then the feature vector is scaled by Equation (6.10). As a result, the scaled feature values are now uniformly distributed between zero and one. A disadvantage of it is that we neglect the information provided by the differences between feature values, because we only look at the order of them. However, the advantage of it is that we know what the scaled feature values will be and, by that, can design the transformation function such that the resulting IFDs have certain properties. We will use both the original feature values and ranked feature values in this thesis.

Smallest and Largest Individual Feature Deviation As we showed, two forecasts have a scaled feature value of zero and one respectively. Those forecasts are connected to the remaining common arguments of the transformation function of Equation (6.5). The two arguments of Ψ are the *smallest deviation*, ψ_{min}, and *largest deviation*, ψ_{max}. As their name suggests, they define the smallest and largest deviation a weight can be assigned, i.e., for example $\omega_i - \dot{\omega}_i \leq \overline{\aleph}_i \in \left[\psi_{min}, \psi_{max}\right] \; \forall \; i = 1, \ldots, N$. The smallest element of the IFD vector is always $min(\boldsymbol{\aleph}) = \psi_{min}$ and the largest element is always $max(\boldsymbol{\aleph}) = \psi_{max}$. For the sake of simplicity, we use the upper IFD $\overline{\aleph}$ in this example

[44] We will discuss how we can impose different properties onto the IFD in Section 6.1.2.

and will do so in the following examples. However, it holds accordingly for the lower IFD $\underline{\aleph}$.

Consequently, there are two forecasts or weights $k, l \in \{1, \ldots, N\}$ from the optimization problem in Equation (6.3) that will always have the following two *upper* individual feature bounds

$$\omega_k - \dot{\omega}_i \leq \aleph_k = \psi_{min}, \tag{6.11}$$

$$\omega_l - \dot{\omega}_i \leq \aleph_l = \psi_{max}. \tag{6.12}$$

The forecast k has the scaled feature value $\tilde{\nu}_k = 0$ and l has $\tilde{\nu}_l = 1$. To put it differently, the forecast with the least favorable feature value is constrained by Equation (6.11). It has the tightest bound around or smallest upper deviation from its prior weight. In contrast, the forecast with the most favorable feature value has is constrained by Equation (6.12), i.e., it hast the least tight bound around or largest upper deviation from its prior weight.

For both the smallest and largest deviation (ψ_{min}, ψ_{max}), we need to define conditions to ensure the feasibility of optimization problem of Equation (6.3). Recall, that we defined $\underline{\aleph}_i, \overline{\aleph}_i \in \mathbb{R}_{\geq 0}$. Hence, for the smallest deviation it has to hold that

$$\psi_{min} \geq 0. \tag{6.13}$$

$\overline{\aleph}_i$ defines how much a weight ω_i can deviate from its prior weight $\dot{\omega}$ upwards, i.e., how much larger it can get. This holds accordingly for $\overline{\aleph}_i$. A non-negative smallest deviation is given by design. Otherwise, if both ψ_{min} of $\underline{\aleph}_i$ and ψ_{min} of $\overline{\aleph}$ are non-positive, it can happen that $\underline{\aleph}_i < 0$ and $\overline{\aleph}_i < 0$. Thereby, ω_i is forced to be larger and smaller than the prior weights simultaneously, see Equation (6.4). This is infeasible.

Negative values of $\aleph_i$ and, by that, negative values for ψ_{min} are only sensible if either the lower or the upper IFB ($\underline{\omega}_i$ or $\overline{\omega}_i$) are used but not both. For only lower IFBs, it implies that some forecasts (at least one) are constrained to have larger values than their prior weights. Similarly, for only upper bounds some (at least one) forecasts with more (the most) unfavorable feature values are forced to have smaller weights than their prior weights. In this thesis we will only consider forecast combination with individual feature bounds that have both lower and upper IFB simultaneously.

To define feasible or sensible values of the largest deviation ψ_{max} is more difficult. First, it depends on how the IFBs or rather the prior weights are designed. Assume we use prior weights that are a feasible solution to the forecast combination problem of Equation (2.22), i.e., they fulfill the unity constraint. In this case, conditions for the largest deviation are that it is non-negative and greater than the smallest deviation, i.e., $\psi_{max} \in [0, \infty)$ and $\psi_{min} \leq \psi_{max}$. For the first condition, again, if $\psi_{max} = 0$ the only feasible solution is the prior weights. There is no limit how large ψ_{max} gets, i.e.,

how big the largest deviation is. The second condition ensures that the transformation function that increases as the feature values increase, i.e., the more favorable a forecast value is, the larger the corresponding forecast can deviate from the prior weights.

If the prior weights do not fulfill the unity constraint it is not possible to determine general values for ψ_{max}, as it depends on the distribution of the scaled feature values $\tilde{\boldsymbol{\nu}}$ in connection with the selected transformation function Ψ. For example, assume that all prior weights are zero, i.e., $\dot{\boldsymbol{\omega}} = \mathbf{0}$. In this case, we can determine a minimal value for ψ_{max} that *can* result in a feasible solution. Let us start with an example where all weights deviate from zero by $1/N$. This can result in a feasible solution. However, this is only the case if all scaled feature value are one, i.e., $\tilde{\boldsymbol{\nu}} = \mathbf{1}$. This is no possible due to the way we scale the feature values. Equations (6.9) and (6.10) are undefined in that case (division by zero). However, if $\nu_i = \nu_j \; \forall \; i, j\{1, \dots N\} \setminus k$ with $\nu_k < \nu_j \; \forall j\{1, \dots, N\} \setminus k$, we can define $\psi_{max} = 1/N-1$. In others words, if all forecasts but one have the same favorable feature value, they all get a scaled feature value of one while the remaining forecast has a scaled feature value of zero. For an arbitrary transformation function it holds by definition that $\Psi(\tilde{\nu}_i = 1, \psi_{min}, \psi_{max}, \dots) = \psi_{max}$. Accordingly, if $\psi_{max} = 1/N-1$, $(N-1)$ weights can deviate from zero by that amount which is a feasible solution.

This example showcases, however, that it depends on the distribution within the scaled feature vector $\tilde{\boldsymbol{\nu}}$ which values of the largest deviation ψ_{max} result in a feasible solution. Additionally, it depends on which transformation function together with its arguments, that we will discuss in Section 6.1.2.

In general, the unity constraint will always be violated if

$$\mathbf{1}'(\dot{\boldsymbol{\omega}} + \overline{\boldsymbol{\aleph}}) < 1, \tag{6.14}$$

i.e., if the sum of prior weights and the upper IFDs is smaller one. [45] We can determine an adjusted upper IFD vector $\overline{\boldsymbol{\aleph}}^*$ to ensure feasibility by

$$\overline{\boldsymbol{\aleph}}^* = \frac{\dot{\boldsymbol{\omega}} + \overline{\boldsymbol{\aleph}}}{\mathbf{1}'(\dot{\boldsymbol{\omega}} + \overline{\boldsymbol{\aleph}})} - \dot{\boldsymbol{\omega}}. \tag{6.15}$$

By that the upper IFDs are scaled such that the weights can deviate by a sufficient amount from the prior weights such that the weights sum to exactly one. As a consequence the solution to the optimization problem of Equation (6.3) is $\boldsymbol{\omega} = \boldsymbol{\aleph}^*$ with $\omega_i \geq 0 \; \forall i \; \in \{1, \dots, N\}$, i.e., all weights are positive and identical to their individual bounds. Note that the adjustment of Equation (6.15) ensure that the ratio between the constraints remains the same, i.e., $\overline{\aleph}_i/\overline{\aleph}_j = \overline{\aleph}_i^*/\overline{\aleph}_j^* \; \forall i, j \in \{1, \dots, N\}$. Although, the

[45] Note the feasibility of the optimization problem, i.e., the fulfillment only depends on the upper deviation $\overline{\aleph}$ we do not need to consider the lower IFDs ($\underline{\aleph}$) because it does not affect feasibility if sum of prior weights and upper bound deviations is smaller one.

desired, predefined largest deviation ψ_{max} is exceeded by this approach, Equation (6.15) enables us to calculate a solution for every distribution of scaled feature vector $\tilde{\boldsymbol{\nu}}$. For example, let $N = 2$ with $\dot{\omega}_1 = \dot{\omega}_2 = 0$ and $\overline{\aleph}_1 = 0.2$ and $\overline{\aleph}_2 = 0.4$. The adjusted values are $\overline{\aleph}_2^* = {}^{0.2}/_{0.6} = {}^{1}/_{3}$ and $\overline{\aleph}_2^* = {}^{0.4}/_{0.6} = {}^{2}/_{3}$ which provides a feasible solution.

This holds accordingly if

$$\mathbf{1}'(\dot{\boldsymbol{\omega}} - \underline{\boldsymbol{\aleph}}) > 1, \tag{6.16}$$

with

$$\underline{\boldsymbol{\aleph}}^* = \dot{\boldsymbol{\omega}} - \frac{\dot{\boldsymbol{\omega}} - \underline{\boldsymbol{\aleph}}}{\mathbf{1}'(\dot{\boldsymbol{\omega}} - \underline{\boldsymbol{\aleph}})}. \tag{6.17}$$

Summary We propose to use individual feature bounds of Equations (6.1) and (6.2). The main idea is that we constrain weights of forecasts that have favorable features values less than weights of forecast with unfavorable feature values. To this end, we introduce individual feature deviations (IFD) $\underline{\aleph}_i$ and $\overline{\aleph}_i$, for each weight $\omega_i \; \forall \; i = 1, \ldots, N$. In conjunction with prior weights $\dot{\omega}_i$ the IFDs determine the IFBs $(\underline{\omega}, \overline{\omega})$, i.e., the interval of feasible values for each weight. Thereby, weights are shrunken towards their prior weights. The IFDs are the output of the transformation function Ψ. The inputs to the transformation functions are the scaled feature values $\tilde{\boldsymbol{\nu}}$ as a result of a function ξ of forecasts together with additional parameters like the smallest and largest deviations ψ_{min} and ψ_{max}.

We believe that our newly proposed method or research area has the potential to inspire future research due to its variability. First, one can use a variety of transformation function that can be defined in such a way that they enforce specific conditions onto the IFD vectors $\underline{\boldsymbol{\aleph}}$ and $\overline{\boldsymbol{\aleph}}$ and by that onto the bounds. We will discuss various transformation functions in Section 6.1.2. Second, one can use different features for different applications that can incorporate new information into the optimization problem. This will be discussed in Section 6.1.3. Lastly, one can use different prior weights, i.e., shrinkage directions, that also inform the forecast combination problem.

6.1.2 Transformation Functions

In this section we discuss the design of a major part of the individualized feature bounds: the transformation function Ψ of Equation (6.5). Again, the main idea behind IFBs is to constrain weights or forecasts more or less depending on how favorable they are with respect to a certain feature. The transformation function maps or transforms a vector of (scaled) feature values into individual feature deviations $\underline{\aleph}_i$ and $\overline{\aleph}_i$. In conjunction with the prior weight $\dot{\omega}_i$, this leads to both an individual lower and upper bound $\underline{\omega}_i, \overline{\omega}_i$ for each weight ω_i, see again Equations (6.1) and (6.2).

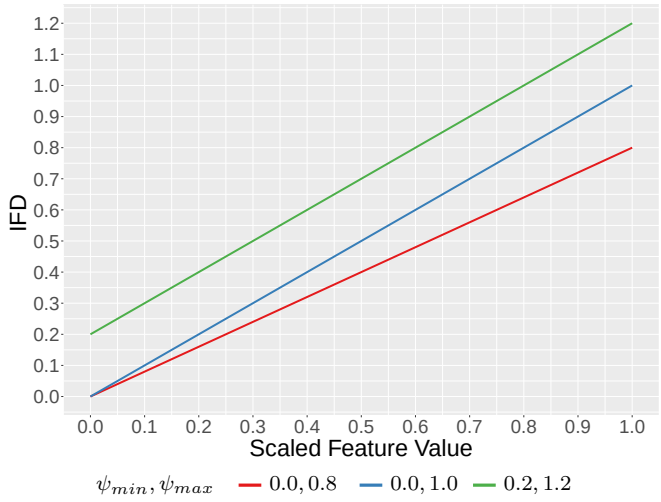


Figure 6.5. Examples of the linear transformation function.

In this thesis we present four transformations functions. We first introduce a linear and second a step-wise transformation function in Sections 6.1.2.1 and 6.1.2.2 respectively. Then we will present an adaptation of an activation function that is used within neurons from artificial neural networks. It will be discussed in Section 6.1.2.3. Additionally, we provide a framework to use a generalized version of the logistic function as a transformation function in Section 6.1.2.4.

6.1.2.1 Linear Function

The first transformation function originates from the well-known basic definition of a *linear* function: $f(x) = b + mx$. Recall, that the scaled feature values are between zero and one and that a larger scaled feature value is favorable. We define the linear transformation function as

$$\Psi\Big(\tilde{\boldsymbol{\nu}}, \psi_{min}, \psi_{max}\Big) = \psi_{min} + \big(\psi_{max} - \psi_{min}\big)\tilde{\boldsymbol{\nu}}. \tag{6.18}$$

Accordingly, Ψ takes the scaled feature values within $\tilde{\boldsymbol{\nu}}$ and transforms or maps them linearly between the smallest deviation ψ_{min} and the largest deviation ψ_{max}. The forecast with the most favorable feature value gets the largest deviation ψ_{max}, i.e., it is the least constraint compared to all other forecast. Similarly, the forecast with the most unfavorable feature value gets the smallest deviation, i.e., it has the tightest constraint in comparison.

Figure 6.5 depicts three possible linear transformation functions Ψ based on Equation (6.18). For both the red and blue line, the smallest deviation is $\psi_{min} = 0$. As a result, the forecast with the least favorable feature value will be constraint to its prior weight. The difference in slope between red and blue is due to the largest possible

bound $\psi_{max} = 0.8$ and 1.0 respectively. When comparing the blue and red line it becomes apparent that, ceteris paribus, for smaller values of ψ_{max}, the weights are more constraint, i.e., deviate less from their prior weights, for all scaled feature values (except for $\tilde{\nu}_i = 0$). By increasing the smallest deviation ψ_{min} to be greater zero, no forecast or weight is forced to be equal to its prior weight. This is depicted by the green line.

A linear transformation function is a simple function that can be used. It basically just re-scales the scaled feature vector $\boldsymbol{\nu}$ onto the interval $[\psi_{min}, \psi_{max}]$. As a result, the ratio between two scaled feature values ν_i and ν_j is identical to the ratio between the resulting IFDs $\aleph_i$ and $\aleph_j$ $\forall i, j \in \{1, \ldots, N\}$.

6.1.2.2 Step Function

The second transition function that we will consider is a *step* function. For a predefined number of steps Υ, the step size is

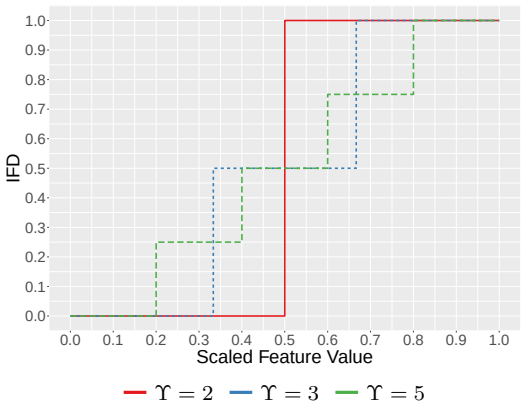
$$\check{\Upsilon} = \frac{\left(\psi_{max} - \psi_{min}\right)}{\Upsilon}. \tag{6.19}$$

For $\Upsilon \in \mathbb{N}_{\geq 1}$ we define the transformation function by

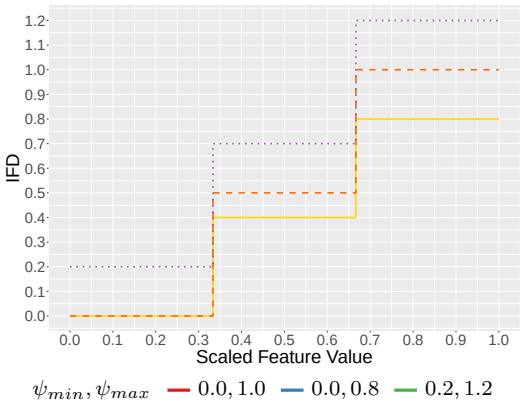
$$\Psi\left(\tilde{\boldsymbol{\nu}}, \psi_{min}, \psi_{max}, \Upsilon\right) = \begin{cases} \psi_{min} & \text{if } \tilde{\nu}_i \leq \frac{1}{\Upsilon} \\ \psi_{min} + \check{\Upsilon} & \text{if } \frac{1}{\Upsilon} < \tilde{\nu}_i \leq \frac{2}{\Upsilon} \\ \psi_{min} + 2\check{\Upsilon} & \text{if } \frac{2}{\Upsilon} < \tilde{\nu}_i \leq \frac{3}{\Upsilon} \\ \vdots & \vdots \\ \psi_{min} + (\Upsilon - 1)\check{\Upsilon} & \text{if } \frac{(\Upsilon-1)}{\Upsilon} < \tilde{\nu}_i \leq \frac{\Upsilon}{\Upsilon} \end{cases} \qquad \forall\ i = 1, \ldots, N. \tag{6.20}$$

This function divides the scaled feature vector $\tilde{\boldsymbol{\nu}} \in [0, 1]$ into Υ different sections or interval, e.g., $\frac{1}{\Upsilon} < \tilde{\nu}_i \leq \frac{2}{\Upsilon}$. Each section has a value that is the output of the transformation function, if the scaled feature value of a forecast is within this section. The first sections output is ψ_{min} for the smallest values in $\tilde{\boldsymbol{\nu}}$. It increases step wise by $\check{\Upsilon}$ until the last section. For the largest values in $\tilde{\boldsymbol{\nu}}$ it becomes ψ_{max}. The only exception is if $\Upsilon = 1$ is used for the step function. In this case the deviations are no longer individual but identical for all forecast with $\aleph_i = \psi_{min}$ $\forall\ i = 1, \ldots, N$, i.e., it is identical to the bounded prior weights approach of Section 5.2. Accordingly, forecast combination with individual features bounds effectively nests forecast combination with bounded weights.

Figure 6.6 depicts examples of the step function defined in Equation (6.20). In both Figures 6.6(a) and 6.6(b) the scaled feature vector $\tilde{\boldsymbol{\nu}}$ is shown on the abscissa. The ordinate depicts the resulting IFDs within $\boldsymbol{\aleph}$. Note that for both figures the different line types are used for visualization purposes only because lines overlap. First, in



(a) Ψ for a different number of steps with $\psi_{min} = 0$ and $\psi_{max} = 1$.



(b) Ψ for different ψ_{min} and ψ_{max} with $\Upsilon = 3$ steps.

Figure 6.6. Examples of the step transformation function Ψ. Different line types are for visualization purposes only.

Figure 6.6(a) different values for the number of steps Υ are used. The red line depicts the step function with two steps, i.e., $\Upsilon = 2$. The two steps are the plateaus or sections for $\tilde{\nu} \leq 1/2$ and $\tilde{\nu} > 1/2$. If a forecast has a scaled feature value smaller or equal to 0.5 its IFD is zero. In contrast, for all forecast with $\nu_i > 0.5$ the IFD is one. If the number of steps is increased to three (blue) or five (green), additional shorter sections are created. By that, the IFD differentiates more between feature values. For example, for both three and five steps a scaled feature value of 0.5 is transformed to 0.5. However, for $\tilde{\nu}_i = 0.65$, the transformation function with $\Upsilon = 5$ (green) results in $\aleph_i = 0.75$ while for $\Upsilon = 3$ (blue) it is still 0.5.

We can use Figure 6.6(a) to think about the step function in terms of quantiles. However, it is important to notice that we use the quantiles of *potential* scaled feature values $\tilde{\boldsymbol{\nu}} \in [0, 1]$ and not quantiles of the actual, realized feature values. This means we use the quantiles of the interval $[0, 1]$. The scaled feature values $\tilde{\nu}$ where the step transformation function transitions from one section to the other are the $1/r{-}1\%$ quantiles $\forall\ r = 2, \ldots, \Upsilon$ of all values between zero and one. For example, if we use $\Upsilon = 5$ (green) in Equation (6.20), we have five sections. The transitions from one section to another are at the $20\%, 40\%, 60\%$ and 80% quantiles of potential scaled feature values. To put it differently, forecasts that are at the bottom 20% of scaled feature values get the corresponding IFD, i.e., zero in case of the green step function. Forecasts with feature values that are within 40% to 60% of the difference between the largest and smallest feature value $\boldsymbol{\nu}$ get an IFD of 0.5.

The blue line of Figure 6.6(a) and the orange line of Figure 6.6(b) are identical with $\Upsilon = 3, \psi_{min} = 0$ and $\psi_{max} = 1$. In Figure 6.6(b) we compare different values for ψ_{min} and ψ_{max} while keeping $\Upsilon = 3$ constant. Compared to the orange line, the yellow line has the same smallest deviation ψ_{min} but a smaller largest deviation $\psi_{max} = 0.8$.

If a forecast i has a scaled feature value less or equal to $1/3$, both orange and yellow transform its value to $\aleph_i = 0$. However, if $\tilde{\nu}_i \in (2/3, 3/3]$ the step function depicted in yellow transforms its value to $\aleph_i = \psi_{max} = 0.8$ while the orange line transforms it to $\aleph_i = \psi_{max} = 1$. The purple line demonstrates an example at which the smallest deviation is 0.2 and the largest deviation is 1.2. Comparing this step transformation against all others demonstrates an opportunity that comes with using a step function. For the purple line all forecasts have IFD $\aleph_i$ of at least 0.2. All other transformation functions in both Figures 6.6(a) and 6.6(b), however, have some forecasts with IFDs of zero.

Accordingly, with the step function if we define $\psi_{min} = 0$, the approach selects variables that are allowed to differ from their prior weights based on their scaled feature values. All forecasts that have a scaled feature value smaller or equal to $1/S$ will have the IFD $\aleph_i = 0$, i.e., their weight will be equal to their prior weight. If all prior weights are zero, this corresponds to a variables selection where some weights are constraint to zero, i.e., the corresponding forecasts are omitted. If we use the unranked scaled feature values, we can not define in advance how many forecasts will be identical to their prior weights. Instead, we define to omit those forecasts that have feature values less or equal to the $1/\Upsilon$ quantile of potential, scaled feature values. Recall, that $\tilde{\boldsymbol{\nu}}$ is zero for the forecast with the least favorable feature value and one for the most favorable feature value. As a consequence, if $\psi_{min} = 0$, at least one forecast is omitted and at most all but one forecasts can be omitted.

For example, consider the step transformation function with $\Upsilon = 5, \psi_{min} = 0$ and $\psi_{max} = 1$ depicted in green in Figure 6.6(a). If $\tilde{\boldsymbol{\nu}} = (0, 0.3, 0.5, 0.8, 1)$ only one forecast is omitted while if instead $\tilde{\boldsymbol{\nu}} = (0, 0.1, 0.15, 0.2, 1)$ all but one forecast are omitted. The case that all but one forecast are omitted only happens if there is a forecast that a significantly more favorable feature values. In contrast, if we use the ranks feature values, we know the number of forecasts within each section of the step function. For example, for a step function with $\Upsilon = 2$ and eleven forecasts, six forecasts get the IFD from section one (ψ_{min}), and five forecasts the IFD from section two (ψ_{max}).

Using a step function opens up many possibilities. Depending on its setup it even can be used a direct variable selection approach. For example, let $\Upsilon = 2$ (red line in Figure 6.6(a)) with $\psi_{max} = M$ and $M \to \infty$ and prior weights of zero. By that we create a scenario where all forecasts that are at the lower half of possible scaled feature values are identical to their prior weights and all others are unconstrained. Similarly, we can let $\psi_{min} = 0.2$ so that all forecast with a feature value less than the median of possible scaled feature values are constrained based on $\aleph_i = 0.2$. All other forecasts are still unconstrained with $\psi_{max} = M$.

There are many more possibilities to create a variety of step functions. In general, the larger the number of steps, ceteris paribus, the more Ψ differentiates between fore-

casts,i.e., their scaled feature values. Simultaneously, the difference between sections, i.e., the step size of Equation (6.19), becomes smaller. If instead S is kept constant and ψ_{max} (ψ_{min}) is increased (decreased), ceteris paribus, the step size $\check{\Upsilon}$ increases (decreases).

6.1.2.3 Generalized RelU

The third transformation function is inspired by the *Rectified Linear Unit (RelU)* function. The RelU function is used as an activation function for a neuron in artificial neural networks, and it is defined as $f(x) = max\{0, x\}$. For negative values of x it is zero and for positive values it increases linearly (for more details see Aggarwal, 2023, Chapter 1, especially pp.10-12). We adapt this function and design the following transformation function:

$$\Psi\left(\tilde{\boldsymbol{\nu}}, \psi_{min}, \psi_{max}, \check{\nu}\right) = \begin{cases} \psi_{min} & \text{if } \tilde{\nu}_i \leq \check{\nu} \\ \psi_{min} + \frac{\psi_{max} - \psi_{min}}{1 - \check{\nu}} \cdot (\tilde{\nu}_i - \check{\nu}) & \text{if } \tilde{\nu}_i > \check{\nu} \end{cases} \tag{6.21}$$

In contrast to the original RelU we additionally use $\check{\nu} \in [0, 1]$. If $\check{\nu} = 0$, the function simplifies to the linear function of Equation (6.18). For any value $\check{\nu} \in (0, 1)$ the function assigns the value of ψ_{min} if $\nu_i \leq \check{\nu}\ \forall\ i \in \{1, \ldots, N\}$. For all $\nu_i > \check{\nu}$ the function linearly interpolates between ψ_{min} and ψ_{max}. Accordingly, for a given threshold $\check{\nu} \in [0, 1]$ all forecasts that are in the bottom $\check{\nu} \cdot 100\%$ of potential feature values have the same IFD $\aleph_i = \psi_{min}$. The top $(1-\check{\nu})\cdot 100\%$ of potential feature values get less constraint the higher their scaled feature value is, i.e., for those forecasts it holds that $\psi_{min} < \aleph_i \leq \psi_{max}$.

The main idea behind the *Generalized RelU (GRelU)* is that forecast with feature values within a certain percentage of potential feature values, i.e., forecast with less favorable feature values, have the same deviation $\aleph_i = \psi_{min}$. At the same time forecast with more favorable feature values are less constraint the larger their feature values are. Similar to the step function, if we chose $\psi_{min} = 0$, the Generalized RelU can perform a variable selection. It selects weights to be equal to their prior weights. If prior weights are zero, forecasts are not considered for forecast combination.

The GRelU is visualized in Figure 6.7. It shows three different designs of the GRelU transformation function, distinguished by color. The scaled feature values are on the abscissa and the IFDs are on the ordinate. Different line types are used for a better visualization if the transformation functions overlap. Both the blue line and green line have $\psi_{min} = 0$ and, by that, forecasts for which $\nu_i \leq \check{\nu} = 0.4$ (blue) or $\tilde{\nu}_i \leq \check{\nu} = 0.6$ (green) $\forall\ i = 1, \ldots, N$ have IFDs of $\aleph_i = 0$, i.e., they are constrained to be identical to their prior weights. All forecast right of the threshold $\check{\nu}$ have bounds greater zero. However, the IFDs differ depending on the largest deviation ψ_{max} and the threshold $\check{\nu}$. Both blue and green have $\psi_{max} = 1$ but for blue $\check{\nu}$ is smaller and, thus, the incline has

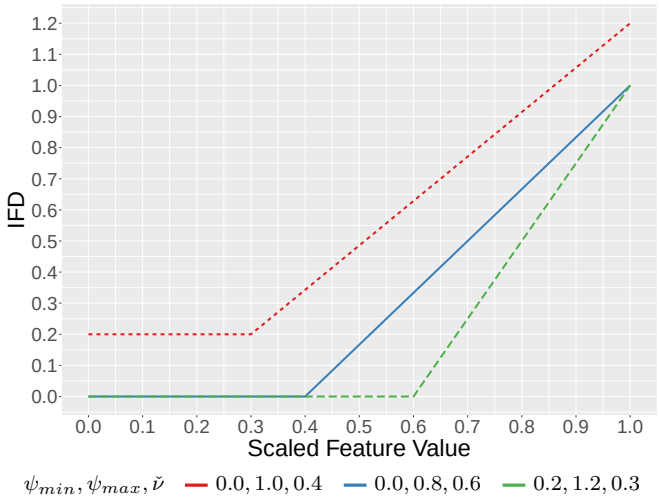


Figure 6.7. Three examples of the GRelU transformation function Ψ. Different line types and colors are used for visualization purposes.

to be steeper to ensure that the larges scaled feature value get the constraint $\aleph_i = \psi_{max}$. The marginal effect of the linear, differentiable part of Equation (6.21) with respect to ν_i is larger the closer $\check{\nu}$ is to one, i.e., the transformation function differentiates more between scaled feature values. While both blue and green perform a variable selection, the transformation function depicted by the red line does not. For all forecast with $\nu_i \leq \check{\nu} = 0.3$ it holds that $\aleph_i = 0.2$, i.e., with regard to the optimization problem from (6.3) it has to hold that $\dot{\omega} - 0.2 \leq \omega_i \leq \dot{\omega} + 0.2$. Forecasts that have a feature value greater the threshold increase linearly to ψ_{max}.

In summary, the GRelU allows the same deviation for a certain percentage of potential scaled feature values. For the remaining one it assigns different deviations based on their scaled feature values. If ranked feature values are used, the percentage values corresponds to the number of forecasts.

6.1.2.4 Generalized Logistic Function

For the last transformation function we want to define a function that is less sensitive to differences in the scaled feature values that are close to the most unfavorable or favorable values. However, it should be more sensitive to changes of the scaled feature values for a certain range of values. To this end, we use a non-linear but differentiable function introduced in Richards (1959). It is a growth function that is a generalization of the logistics function. It is defined as

$$f(x) = \frac{\psi_{max}}{(1 + \phi_1 e^{-\phi_2 x})^{1/\phi_3}}, \tag{6.22}$$

with the constants $\phi_1 \in \mathbb{R}_{\geq 0}$, $\phi_2 \in \mathbb{R}$ and $\phi_3 \in \mathbb{R}_{>0}$. For $\phi_1, \phi_2, \phi_3, \psi_{max} = 1$ it is the commonly known logistic function (Aggarwal, 2023, pp. 81-82). The limits of the generalized logistic function of Equation (6.22) are $\lim_{x \to +\infty} f(x) = \psi_{max}$ and $\lim_{x \to -\infty} f(x) = 0$ (Causton, 1969; Richards, 1959).

The generalized logistics function can look quite different based on the values of ϕ_1,ϕ_2 and ϕ_3. Before we take a closer look at those parameters, we first adapt the generalization of the logistics function for our transformation function. Previously, when we considered functions for the transformation function $\Psi\big(\boldsymbol{\nu}, \psi_{min}, \psi_{max}, ...\big)$, we defined the smallest and largest deviation $\aleph_i$ by ψ_{min} and ψ_{max}. To make this possible with the generalized logistics function, we can linearly scale the function of Equation (6.22) onto the interval $[\psi_{min}, \psi_{max}]$ by

$$\tilde{f}(x) = \psi_{min} + (1 - {}^{\psi_{min}}/_{\psi_{max}}) f(x). \tag{6.23}$$

As a result, the lower limit of $\tilde{f}(x)$ is ψ_{min}, i.e.,

$$\lim_{x \to -\infty} \tilde{f}(x) = \psi_{min}, \tag{6.24}$$

but the upper limit is

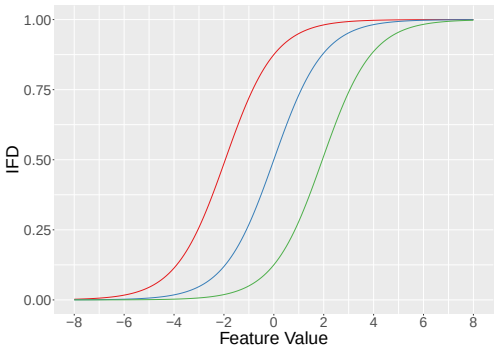
$$\lim_{x \to \infty} \tilde{f}(x) = \psi_{max}. \tag{6.25}$$

We define the transformation function Ψ to be

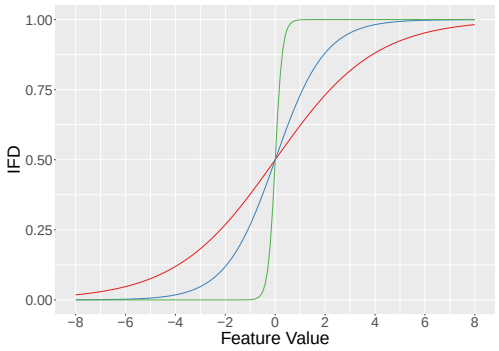
$$\Psi\Big(\tilde{\nu}_i, \psi_{min}, \psi_{max}, \phi_1, \phi_2, \phi_3\Big) = \psi_{min} + \frac{\psi_{max} - \psi_{min}}{(1 + \phi_1 e^{-\phi_2 \tilde{\nu}_i})^{(1/\phi_3)}}. \tag{6.26}$$

This transformation function constrains the deviation of weights from their prior weights similarly if the corresponding forecasts have comparable, unfavorable scaled feature values. This holds accordingly for forecasts with comparable, favorable scaled feature values. This result from the fact that the function flattens both for smaller and larger values of the scaled feature values. In the middle of input values into the generalized logistics function it is a lot more steep, i.e., small difference in the scaled feature values result in more noticeable differences in $\aleph_i$. All this depends on the parameter values of ϕ_1,ϕ_2 and ϕ_3.

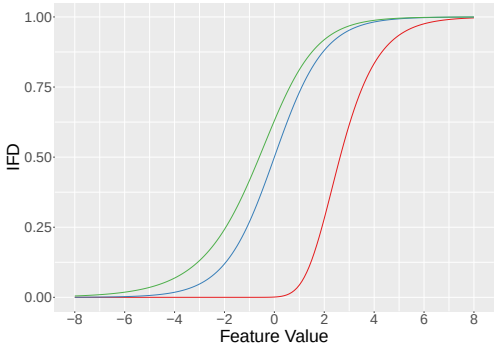
Variations of the Transformation Function Figure 6.8 depicts different variations of $\Psi\big(\boldsymbol{\nu}, \psi_{min}, \psi_{max}, \phi_1, \phi_2, \phi_3\big)$ defined in Equation (6.26). Note that the abscissa represents ν (not scaled) and also includes negative values. The IFD is given on the ordinate. For now, we use the not scaled feature values for the sake of simplicity. As a baseline throughout Figure 6.8 we use a smallest deviation of $\psi_{min} = 0$, a largest



(a) Shift Parameters $\phi_1 = 1/7$ (red), $\phi_1 = 1$ (blue) and $\phi_1 = 7$ (green).



(b) Growth Parameters $\phi_2 = 0.5$ (red), $\phi_2 = 1$ (blue) and $\phi_2 = 7$ (green).



(c) Shift Parameters $\phi_3 = 0.1$ (red), $\phi_3 = 1$ (blue) and $\phi_3 = 1.5$ (green).

Figure 6.8. Examples for the generalized logistic function of Equation (6.22) for different values for ϕ_1,ϕ_2 and ϕ_3, ceteris paribus. For the basic set up we define $\phi_1 = \phi_2 = \phi_3 = 1$.

deviation of $\psi_{min} = 1$ and define $\phi_1 = \phi_2 = \phi_3 = 1$. In Figures 6.8(a) to 6.8(c) we vary the parameters ϕ_1,ϕ_2 and ϕ_3 ceteris paribus.

Figure 6.8(a) emphasizes the effect of the parameter ϕ_1. With $\phi_1 = 1$ it holds that $\Psi(0, 0, 1, 0.5, 1, 1) = 0.5$. If ϕ_1 deviates from one, it shifts the whole function to the right if $\phi_1 > 1$, and to the left if $\phi_1 < 1$. Then for $\phi_1 > 1$ it holds that $\Psi(0, 0, 1, 0.5, 1, 1) < 0.5$ and for $\phi_1 < 1$ $\Psi(0, 0, 1, 0.5, 1, 1) > 0.5$. The former case is depicted in Figure 6.8(a) by the green line with $\phi_1 = 7$ and the latter case is shown by the red line with $\phi_1 = 1/7$. The blue line is the baseline with $\phi_1 = 1$. Let us consider the implication for the IFD. We can design a transformation function that constrains forecasts with feature values within the bottom $q\%$ of potential feature values similarly close to ψ_{min}. In the same sense, we can constrain forecasts with feature values within the top $p\%$ of potential feature values close to ψ_{max}. By the choice of ϕ_1, we either increase q or p. For example, if we want to increase the range of feature values that have an IFD close to ψ_{max}, we can decrease ϕ_1 and by that increase p, ceteris paribus. See again the red line in Figure 6.8(a). In that case, all forecast with a feature value less than roughly -6, i.e., about $1/8 \cdot 100\%$ have

an IFD close to zero while forecasts with $\nu_i > 2$, i.e., about $^6/_{16} \cdot 100\%$, have $\aleph_i$ close to ψ_{max}. Similarly, if we want to increase the range of potential feature values that have a constraint close to ψ_{min} instead, we shift Ψ to the right by increasing ϕ_1. As a result, only about the top $^1/_8 \cdot 100\%$ of feature values map to IFDs close to ψ_{max}. In contrast, the bottom $^6/_{16} \cdot 100\%$ of feature values are close to ψ_{min}.

The second parameter, ϕ_2, determines the growth rate, i.e., how steep or shallow the function growth from ψ_{min} to ψ_{max}. This is depicted Figure 6.8(b) exemplary. The blue line shows the function values for $\phi_2 = 1$. If ϕ_2 increases, the growth becomes much steeper as showcased by the green line with $\phi_2 = 7$. If, instead ϕ_2 decreases, the function flattens as depicted by the red line ($\phi_2 = 0.5$). In the context of the IFD $\aleph_i$ we can, again, increases the range of feature values that are either close to ψ_{min} or ψ_{max}. At first, this is similar to the effect of parameter ϕ_1. However, by changing ϕ_2 we de- or increase p and q simultaneously. Basically, the larger ϕ_2 is, the more feature values both on the right and left side of the inflection point are close to ψ_{min} or ψ_{max}, ceteris paribus. Thereby, less feature values result in a bound in-between the smallest and largest deviation. This can be seen easily when comparing the red, blue and green line. To put it differently, the larger ϕ_2 gets the closer this transformation function is to a step function with $\Upsilon = 2$, see Section 6.1.2.2.

Lastly, the parameter ϕ_3 influences the position of the inflection point. For $\phi_3 = 1$ the inflection point is at $\nu' = 0$ and $\Psi(\nu', 0, 1, 1, 1, 1) = 0.5$. If $\phi_3 > 1$, the inflection point $\nu' < 0$ and $\Psi(\nu', 0, 1, 1, 1, > 1) > 0.5$. Accordingly, if $\phi_3 < 1$, the inflection point $\nu' > 0$ and $\Psi(\nu', 0, 1, 1, 1, > 1) < 0.5$. In Figure 6.8(c) the blue line depicts the baseline with $\phi_3 = 1$. The green line shows the function for a higher value of $\phi_3 = 1.5$. The inflection point is higher on the ordinate and more left on the abscissa compared to the blue line. Similarly, for a smaller value of $\phi_3 = 0.1$, the inflection point is lower (Causton, 1969; Richards, 1959). The effect of parameter ϕ_3 in context of $\aleph_i$ is similar to parameter ϕ_1, however only to an extent. If we choose $\phi_3 < 1$ the inflection point is lower, see for example the red line. By that, we increase the steepness at first. For forecasts with smaller feature values ν_i, small changes in ν_i have a larger effect. However, the function than flattens for larger values of ν_i. This means that the difference in the feature values of two forecasts has to be larger to result in the same increase in $\aleph$. For example for $\phi_3 = 0.1$, scalded feature values $\nu_i \in [0, 2]$ are mapped to roughly $\aleph_i \in [0.001, 0.281]$. In contrast, for $\nu_i \in [2, 4]$ the deviations are about $\aleph_i \in [0.281, 0.834]$, i.e., a much larger difference. Increasing ϕ_3 has a similar but inverted effect.

Of course, on can vary all parameters ϕ_1,ϕ_2 and ϕ_3 at once. Accordingly, there are many designs of the generalized logistics function that can be used as a transformation function to determine how weights are constraint around their prior weights.

Procedure for using Ψ Using the generalized logistics function as a transformation function Ψ is not straightforward but comes with its difficulties. They are driven by the asymptotic behavior and the amount of possible function defined by varying ϕ_1, ϕ_2 and ϕ_3. In Figure 6.8 we did not scale the feature values due to this. Of course, all functions are valid for all values of $\nu_i \in \mathbb{R}$. However, trying to design the transformation function with ϕ_1, ϕ_2 and ϕ_3 such that it behaves in a desired way for a *specific range* of feature values, like $\tilde{\boldsymbol{\nu}}$, is quite difficult and arbitrary. Although, we used the scaled feature value to ensure comparability between different transformation previously, we will not do this for the generalized logistics function due to the mentioned difficulties.

However, it is also not straightforward onto which interval one has to scale $\boldsymbol{\nu}$ for the generalized logistics curve. For example, the IFDs that result from the transformation functions depicted in Figure 6.8(b) differ drastically depending on whether we have $\nu_i \in [-1, 1]$ or if we instead have $\nu_i \in [-6, 6]$. In the latter case, the majority of potential feature values are close to zero or one for the green line, while for the blue line this is true for a smaller portion. In contrast, for $\nu_i \in [-1, 1]$ most values of the blue line are between about $[0.25, 0.75]$ rendering our desired smallest and largest deviation ψ_{min} and ψ_{max} pointless. Therefore, when using the generalized logistics function, we do not scale the feature vector between zero and one as we did previously for the linear, step and GRelU function.

Instead, for a given function Ψ, we scale the feature vector such that it has a desired output for certain inputs. To this end, we set

$$\Psi\Big(\nu_i, \psi_{min}, \psi_{max}, \phi_1, \phi_2, \phi_3\Big) = \aleph', \tag{6.27}$$

and solve for ν_i resulting in the inverse function of Ψ, i.e.,

$$\Psi^{-1}\Big(\aleph', \psi_{min}, \psi_{max}, \phi_1, \phi_2, \phi_3\Big) = \frac{1}{\phi_2} \ln\left(\frac{\phi_1}{\left(\frac{\psi_{min}-\psi_{max}}{\psi_{min}-\aleph'}\right)^{\phi_3} - 1}\right). \tag{6.28}$$

While Ψ maps a certain feature value ν_i to a bound $\aleph_i$, the inverse function Ψ^{-1} maps a given bound $\aleph'$ to the corresponding feature value ν'. Due to the asymptotic nature of the logistic function it has to hold that $\aleph' \in (\psi_{min}, \psi_{max})$, i.e., $\aleph'$ has to be in-between but not equal to the limits of the generalized logistics functions of Equations (6.24) and (6.25). However, we want to ensure that the smallest and largest feature values are mapped to the actual desired values of ψ_{min}, ψ_{max}. To this end, we define

$$\nu_{min} = \Psi^{-1}\Big(\aleph' = \psi_{min} + \zeta, \dots\Big), \tag{6.29}$$

$$\nu_{max} = \Psi^{-1}\Big(\aleph' = \psi_{max} - \zeta, \dots\Big), \tag{6.30}$$

with ζ being an arbitrary small number greater zero. Note that now the lower and upper limit of this function are $\psi_{max} - \zeta$ and $\psi_{max} + \zeta$. With regard to the scaling of the feature vector, ν_{min} is the smallest and the ν_{max} is the largest value of the interval onto we scale $\boldsymbol{\nu}$, i.e., it holds that $\check{\nu}_i \in [\nu_{min}, \nu_{max}] \; \forall \; i = 1 \ldots, N$. If a larger feature value is more favorable we use

$$\check{\boldsymbol{\nu}} = \frac{\boldsymbol{\nu} - \nu_{min}}{\nu_{max} - \nu_{min}}, \tag{6.31}$$

and if the smallest feature value is most favorable we instead use

$$\check{\boldsymbol{\nu}} = \frac{\nu_{max} - \boldsymbol{\nu}}{\nu_{max} - \nu_{min}}. \tag{6.32}$$

As a result, for any given specification of $\Psi\big(\nu_i, \psi_{min}, \psi_{max}, \phi_1, \phi_2, \phi_3\big)$, we create a scaled feature vector $\check{\boldsymbol{\nu}}$ such that the forecast with the least favorable feature value has an IFD of $\aleph_i = \psi_{min} + \zeta$ and the forecast which has the most favorable feature value has $\aleph_i = \psi_{max} - \zeta$. Henceforth, we use $\zeta = 10^{-6}$.

With this design we can achieve the actually desired smallest and largest deviations ψ_{min} and ψ_{max} by a slight modification of Equation (6.26). We simply use the transformation function $\Psi\Big(\nu_i, \check{\psi}_{min}, \check{\psi}_{max}, \phi_1, \phi_2, \phi_3\Big)$ with

$$\check{\psi}_{min} = \psi_{min} - \zeta, \tag{6.33}$$

$$\check{\psi}_{max} = \psi_{max} + \zeta. \tag{6.34}$$

In summary, to properly design and use the generalized logistic function as a transformation function, we need to adjusted the values for the smallest and largest deviation, $\check{\psi}_{min}$ and $\check{\psi}_{max}$ and scale the feature vector accordingly to $\check{\nu}_i \in [\nu_{min}, \nu_{max}]$ $\forall \; i = 1 \ldots, N$. As a result, the transformation function Ψ maps the smallest and largest feature value to the actually desired IFDs, i.e., $min(\boldsymbol{\aleph}) = \psi_{min}$ and $max(\boldsymbol{\aleph}) = \psi_{max}$ respectively.

Concluding Remarks Due to the many different possible parameter combinations of ϕ_1, ϕ_2 and ϕ_3, using the generalized logistics functions opens up a variety of different transformation functions with different properties. This, however, comes with a problem. First, finding the right combination of ϕ_1, ϕ_2 and ϕ_3 such that the transformation function has the desired properties on the desired range of inputs. Second, it is not straightforward to ensure that the smallest and largest deviation are actually part of the IFB $\boldsymbol{\aleph}$. We solved this by scaling the feature values differently and use adjusted smallest and largest deviations $(\check{\psi}_{min}, \check{\psi}_{max})$.

There is, however, a disadvantage when using the generalized logistic function. The number of parameters ($\check{\psi}_{min}$, $\check{\psi}_{max}$, ϕ_1,ϕ_2 and ϕ_3) leads to a substantial increases in

the number of scenarios to evaluate if cross-validation is used. For each additional candidate value of any parameter that has be considered (for example ϕ_1), the number of additional scenarios increases by the product of all candidate values for all other parameters. As a consequence, a large number candidate values for each parameter value is impractical or even infeasible. Therefore, the choice of candidate values for each parameter has to be well curated and, by that, it depends on prior believes.

The generalized logistics function as a transformation function enables us to create a variety of functions. Generally speaking, they are less sensitive to differences in the feature values of forecasts close to the most unfavorable or favorable feature value. Simultaneously, they are more sensitive to small changes in the feature values for forecasts that are closer to the middle of potential feature values. As a result, forecasts with comparable, unfavorable feature values get similar IFDs close to $\check{\psi}_{min}$. This applies accordingly if the feature values of forecasts have comparable, favorable feature values with similar IFDs close to $\check{\psi}_{max}$. For forecasts that have feature values neither close to the most unfavorable nor favorable possible feature values the IFDs are more dissimilar.

6.1.3 Forecast Features

For our proposed approach of forecast combination with individualized feature bounds, we assumed that there is a feature vector $\boldsymbol{\nu}$ that contains feature values ν_i for all forecasts $i = 1, \ldots, N$. We defined $\nu_i = \xi(\boldsymbol{y}_i)$ in Equation (6.6), i.e., the feature value is calculated based on the forecasts themselves. In general, one can use different functions ξ to create the feature vector. For example, one could also consider the in-sample forecast as a time series and calculate time series characteristics for it. However, within this thesis and section we focus on three functions of ξ. To this end, we first consider forecast accuracy measures in Section 6.1.3.1. Then, we take a look at forecast diversity in Section 6.1.3.2. Lastly, in Section 6.1.3.3 we combine both accuracy and diversity to a new feature.

6.1.3.1 Forecast Accuracy

For the evaluation of forecast accuracy or performance a variety of measures can be used, each with its advantageous and disadvantageous. They are based on the forecast error first mentioned in Chapter 2. Recall that the forecast error for any forecast $i = 1 \ldots, N$ is defined as $\varepsilon_{t,i} = y_t - \hat{y}_{t,i}$ for each point in time $t = 1, \ldots, \tau$. Hyndman

and Koehler (2006) provide an extensive overview of popular forecast accuracy measures that includes, inter alia, the[46]

$$\text{Mean Absolute Error (MAE)} = \text{mean}(|\varepsilon_{t,i}|), \tag{6.35}$$

$$\text{Mean Squared Error (MSE)} = \text{mean}(\varepsilon_{t,i}^2), \tag{6.36}$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \text{mean}\left(100\frac{|\varepsilon_{t,i}|}{y_t}\right). \tag{6.37}$$

The MSE is a commonly used measure that squares the forecast errors and, by that, is more sensitive to outliers. In contrast, the MAE weights each forecast error equally, no matter the size of the deviation. Both the MAE and MSE are *scale-dependent accuracy measures*, i.e., they have the same scale and, in case of the MAE, unit as the data. Scale-dependent measures can be used for comparing the accuracy of different models based on the same data. However, as soon as times series or data with different scales are used, scale-dependent measures are not sensible for comparison.

In that case, accuracy measures based on *percentage errors* can be used, e.g., the MAPE. However, if the time series consists of small values, i.e., the absolute error is divided by a small number y_t, its distribution becomes skewed. Even more problematic, the MAPE can not be used if the time series or data includes zero values as it becomes undefined. An example where this can become a problem is for intermittent demand (Hyndman & Koehler, 2006; Kim & Kim, 2016).

Accuracy measures also can be based on *relative errors*. To this end, each forecast error $\varepsilon_{t,i}$ is divided by another forecast error from a benchmark method. Based on the same idea one can use *relative accuracy measures*. In that case the forecast accuracy measure, for example the MSE, of one method is divided by the same forecast accuracy measure of another benchmark method. Both types of accuracy measures are scale-independent.

Lastly, Hyndman and Koehler (2006) proposed the *mean absolute scaled error (MASE)* which scales the forecast error by the in-sample MAE from the naive forecast, i.e., $\hat{y}_t = y_{t-1}$. It is a scale-independent accuracy measure, and it overcomes the problems of *relative errors* and *relative accuracy measures*, like for example the necessity of multiple out-of-sample forecast in order to be able to compute them. For more details on forecast accuracy measures see Hyndman and Koehler (2006) as well as Kim and Kim (2016).

Although there multiple potential forecast accuracy measure that can be used, we will, henceforth, focus on the MSE. It is an accuracy measure that is a commonly used and well-known. Also, the disadvantage of incomparability between different time series is not a factor. We use the MSE to define bounds for our forecast combination problem,

[46] First, note that instead of the mean, the median is also used. Second, we use the following different style of notation for the accuracy measures to be in line with the mentioned work by Hyndman and Koehler (2006).

i.e., we compare the MSE of a forecast only with other MSE values of forecasts of the same time series. Moreover, the very objective of the original forecast combination problem of Equation (2.22) is to minimize error variance of the combined forecast, i.e., its in-sample MSE loss. Lastly, Elliott and Timmermann (2004) showed that the optimal weights under the MSE loss are optimal for a variety of loss functions as long the forecast errors follow an elliptically symmetric distribution (Elliott and Timmermann 2016, pp. 313-315; Elliott and Timmermann 2004).

6.1.3.2 Forecast Diversity

Beside the forecast accuracy, there is another measure that is considered with increasing interest when it comes to forecast combination: *diversity* (see e.g., J. M. Bates & Granger, 1969; Kang, Cao, Petropoulos, & Li, 2022; Thomson et al., 2019; Wang et al., 2023). To emphasize the importance and usefulness of diversity, we make a short excursion into the literature.

Diversity in Combinations In Krogh and Vedelsby (1994) diversity appears for ensembles of neural networks. This means, that different neural networks are trained. Their forecasts are then combined using, for example, the simple average in case of a regression problem. Krogh and Vedelsby (1994) argue, that members of an ensemble or forecasts should have as much disagreement or diversity as possible among them to provide better results. At about the same time in the context of more traditional forecast combination, Batchelor and Dua (1995) found that combinations that consist of more diverse forecasts improve the accuracy, especially if number of forecasts is small. The usefulness of diverse forecast was also identified by Lichtendahl and Winkler (2020) in the context of the M4 Competition (Makridakis et al., 2018, 2020). Recently, Kang et al. (2022) introduced an approach that is based upon FFORMA. Recall, that FFORMA is a meta-learning approach that uses characteristics of historical time series data to determine weights for a given pool of forecasts, and it placed second in the M4 competition. In contrast, the approach by Kang et al. (2022) uses diversity measures and, by that, achieves a comparable accuracy to FFORMA (Kang et al., 2022; Wang et al., 2023; Wang, Kang, Petropoulos, & Li, 2022). Lastly, in Wang, Kang, Petropoulos, and Li (2022) they propose a trimming algorithm that chooses a subset of forecasts based on both accuracy and diversity.

One way to measure the diversity of a set of forecasts is the forecast error correlation. A smaller correlation implies less similarity, i.e., more diversity (Lichtendahl & Winkler, 2020). Atiya (2020) analyzed the variance of the combined forecast, i.e., $V(\hat{y}_c) = E[(\hat{y}_c - E(\hat{y}_c))^2]$, in the presence of a non-negativity constraint for each weight, i.e.,

PW presented in Equation (2.30). They showed theoretically that for the variance of the combined forecast with non-negative weights it holds that

$$V(\hat{y}_c) \leq \sum_{i=1}^{N} \omega_i \sigma_i^2 - 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \omega_i \omega_j (1 - \rho_{ij}) \sigma_i \sigma_j. \tag{6.38}$$

The fist part of the right-hand-side is the weighted average of the individual forecasts variances, i.e., their forecast accuracy. The second term is always non-negative. First, a non-negativity constraint is imposed, i.e., weights are non-negative. Second, a correlation is by definition between negative and positive one, i.e., $(1 - \rho_{ij})$ is always non-negative. Third, the individual standard deviations are, by definition, positive. Accordingly, the smaller the correlation among forecasts, the larger the reduction in the variance of the combined forecast ceteris paribus (Atiya, 2020; Wang, Kang, Petropoulos, & Li, 2022).

Although, the correlation is a measurement for the diversity among forecasts, there are potential problems depending on the use case. For example, if we want to use the forecasts correlations within forecast combination approach with individual feature bounds, we would have to average the correlations of forecast i with all other forecasts $j = 1, \ldots, N$ with $j \neq i$ to come up with a single feature value. However, correlations are not additive and, thus, this is not sensible (Wang, Kang, Petropoulos, & Li, 2022).

Another way to address the diversity of two forecasts was introduced by Thomson et al. (2019) who define a measure of coherence. It is a measurement that indicates the diversity of two forecasts. The *Mean Squared Error for Coherence (MSEC)* of two forecasts $i, j \in \{1 \ldots, N\}$ with $i \neq j$ over the in-sample training set $t = 1, \ldots, \tau$ is given by

$$\text{MSEC}_{i,j} = \frac{1}{\tau} \sum_{t=1}^{\tau} (\hat{y}_{i,t} - \hat{y}_{j,t})^2. \tag{6.39}$$

If two forecasts make the same predictions, the MSEC is zero. Otherwise, the larger the MSEC, the more diverse two forecasts.

The same measurement is considered by Kang et al. (2022) who showed that the MSE of a combined forecast can be written as the following MSE decomposition:[47]

$$\text{MSE}_c = \frac{1}{\tau} \sum_{t=1}^{\tau} \left[\sum_{i=1}^{N} \omega_i \hat{y}_{i,t} - y_t \right]^2, \tag{6.40}$$

$$= \frac{1}{\tau} \sum_{t=1}^{\tau} \left[\sum_{i=1}^{N} \omega_i (\hat{y}_{i,t} - y_t)^2 - \sum_{i=1}^{N-1} \sum_{j>i}^{N} \omega_i \omega_j (\hat{y}_{i,t} - \hat{y}_{j,t})^2 \right], \tag{6.41}$$

[47] Note that Thomson et al. (2019) have a similar formula under the assumption of equal weights.

$$= \sum_{i=1}^{N} \omega_i \text{MSE}_i - \sum_{i=1}^{N-1} \sum_{j>i}^{N} \omega_i \omega_j \text{MSEC}_{i,j}. \tag{6.42}$$

Accordingly, it consists of an accuracy term and a diversity term. The former is the weighted sum of the individual forecast accuracies. The latter term is also weighted by the individual weights, but it is based on a diversity or coherence measure. If all forecasts are perfectly correlated, i.e., there is no diversity because they have identical forecasts, the MSE of the combined forecast is identical to the weighted sum of the individual MSE values. If however, there is a certain diversity within the set of forecasts, the combined MSE becomes smaller than the weighted average of the individual MSE values (Kang et al., 2022; Wang, Kang, Petropoulos, & Li, 2022). Importantly, in comparison to the theoretical results from Atiya (2020) shown in Equation (6.38), this is not limited to the forecast combination problem with a non-negativity constraint.

The results from Kang et al. (2022) shows that a larger diversity within a set of forecast, i.e., a larger MSEC, leads to a better forecast accuracy of the combined forecasts. However, we want to emphasize that it holds ceteris paribus. Of course, one can not simply include forecast with an extremely large difference in their predictions to increase the MSEC. In that case the individual forecast accuracies decrease and, by that, the overall accuracy can decrease. It is a trade-off between the individual forecast accuracy and the diversity of forecasts, i.e., the first and second part of Equation (6.42).

Diversity as a Feature Based on the previous insights, we include diversity as a feature.

To this end, we use the MSEC of Equation (6.39). One might argue that within forecast combination we use the forecast errors to determine the weights and, therefore, the diversity of forecast *errors* (ε) has to be used instead of the diversity of forecasts ($\hat{y}$). However, it holds that

$$\text{MSEC}_{i,j} = \frac{1}{\tau} \sum_{t=1}^{\tau} (\hat{\varepsilon}_{i,t} - \hat{\varepsilon}_{j,t})^2, \tag{6.43}$$

$$= \frac{1}{\tau} \sum_{t=1}^{\tau} (y_t - \hat{y}_{i,t} - y_t + \hat{y}_{j,t})^2, \tag{6.44}$$

$$= \frac{1}{\tau} \sum_{t=1}^{\tau} (\hat{y}_{i,t} - \hat{y}_{j,t})^2. \tag{6.45}$$

If forecasts errors are used, the actual values y_t cancel out. Therefore, in case of the MSEC forecasts or forecast error lead to the same result.

To use the MSEC as feature values ν_i, we calculate it for each forecast i with each other forecast $j = 1, \ldots, N$ with $j \neq i$ and then average it, i.e.,

$$\text{AvgMSEC}_i = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \text{MSEC}_{i,j} \qquad i = 1, \ldots, N. \tag{6.46}$$

With $\nu_i = \text{AvgMSEC}_i$ we have a measurement of how diverse forecast i is on average in comparison to all other candidate forecasts. In the context of forecast combination with IFBs: a more diverse forecast will be less constrained.

6.1.3.3 Accuracy and Diversity

From the MSE decomposition shown in Equation (6.42) it became clear that the forecast accuracy of the combined forecast is influenced by both forecast accuracy of all individual forecast and their diversity.

Weights within the forecast combination framework are determined by the same logic. The objective is to minimize the error variance of the combined forecast. It is a function of both the weighted individual error variances of the forecast, i.e., a measure of accuracy, and the weighted covariances, a measure of diversity. For an example, recall the error variance of the two forecast scenario in Equations (2.17) and (2.18). Accordingly, a problem is that neither the individuals MSE nor their AvgMSEC capture all factors that influence the forecast accuracy of the combined forecast. We will nevertheless use them and argue that imposing constraint based on either the MSE or AvgMSEC sets more emphasis on either the forecast accuracy or diversity.

However, we also want to create a feature that incorporates both the MSE and AvgMSEC, i.e., accuracy and diversity. To this end, we simply calculate the scaled feature values ν_i for all forecasts first based on the MSE, $\tilde{\nu}_i^{MSE}$, and second based on the AvgMSEC, $\tilde{\nu}_i^{AvgMSEC}$ such that a feature value of one represent the most favorable feature value. With respect to the MSE it is the smallest MSE of all forecasts, i.e., the most accurate forecast. With respect to AvgMSEC it the largest value, i.e., the most diverse forecast (on average).

For each forecast $i = 1 \ldots, N$ we define

$$\nu_i^{AccDiv} = \frac{\tilde{\nu}_i^{MSE} + \tilde{\nu}_i^{AvgMSEC}}{2}, \tag{6.47}$$

which is the average of $\tilde{\nu}_i^{MSE}$ and $\tilde{\nu}_i^{AvgMSEC}$. Lastly, we again scale ν_i^{AccDiv}. As a result, we have the scaled feature values $\tilde{\nu}_i^{AccDiv}$ that are defined between zero and one. They are a measure of both accuracy and diversity. It works accordingly, if ranks of feature values are used.

Another possible measure for accuracy and diversity could be decomposed MSE of Equation (6.42). However, it depends on the weighted MSE and MSEC of a whole pool of forecast. In this thesis, we will use ν^{AccDiv} of Equation (6.47). Nevertheless, we provide additional options within Appendix C to inspire future research to evaluate other feature values that simultaneously consider accuracy and diversity.

6.1.4 Summary and Discussion

We propose a way to constraint and thereby combined forecasts: individual feature bounds. Forecasts weights are constrained individually based on feature values or characteristics of them.

The IFB restricts each weights to be within a certain interval. These intervals are determined as deviations from prior weights. To this end, we use the individual feature deviation IFD $\aleph$. To determine the IFDs, we use a scaled feature value for each forecast. It is then transformed by a transformation function Ψ (Section 6.1) into the IFD. For the transformation function Ψ we can use multiple functions. We introduced a linear function (Section 6.1.2.1), a step function (Section 6.1.2.2), a generalized version of the RelU function (Section 6.1.2.3) and a generalized version of the logistic function (Section 6.1.2.4). For all transformation functions we can define the smallest and largest deviation that they are supposed to have. Moreover, additional parameters have to be defined that determine the appearance or shape of the transformation function. Thereby, we can impose certain behaviors like, for example, a form of variable selection or control the sensitivity of Ψ for different ranges of potential feature values (Section 6.1.2). There are many transformation functions that can be designed that way. Therefore, we need to pick which transformation function and candidate values for their input parameters we want to consider. Otherwise, it becomes impractical to calculate and evaluate result for each combination of transformation function and candidate inputs parameters.

Therefore, in this thesis we only consider the *step* transformation function with $\Upsilon = 2$ and *GRelU* with $\check{\nu} = 0.5$ for the transformation functions. To understand why we choose the step function, recall that for the BW we used universal lower and upper bounds. The step function with $\Upsilon = 2$ basically, enables us to analyze a setup where we have two different lower and upper bounds for a subset of forecasts. Additionally, with the step function we can perform a variable selection to the prior weights, including zero. We decided to use the GRelU, because, it can also perform the same variable selection. Additionally, by that we can impose a common bound for a set of forecasts and then constrain the resulting forecasts based on their feature values linearly. In other words, we extend the step function with $\Upsilon = 2$ by replacing the second universal bound by a linear increasing one. For *GRelU* we use the threshold $\check{\nu} = 0.5$, i.e., for ranked feature

values half of the forecast have a common bound while the other forecasts with more favorable feature values are constraint individually.

For both the step and GRelU transformation function, we use the actual feature values and ranked feature values. For the feature values we use both the MSE and AvgMSEC, i.e., measures for accuracy and diversity. Additionally, we consider the feature we proposed Equation (6.47), that considers both accuracy and diversity together.

Henceforth, we will refer to the general idea of the presented methods as Individual Feature Bounds (IFB). This includes both with and without prior weights. If we use the Individual Feature Bounds without prior weights we refer to it as IFB($\cdot$). If prior weights are used, we denote that by IFB($\dot{\omega}$). If we consider a particular prior weights, we replace $\dot{\omega}$ with the notation for that prior weight, e.g., IFB($\boldsymbol{\omega}^{E2}$).

6.2 Application: Simulation Study

In this section we analyze and compare the forecast accuracy of the forecast combination methods discussed in this chapter. To this end, we use the simulation study presented in Section 3.2. For a brief summary of scenarios and, in particular, forecast error correlation matrices see Sections 3.2.2.2 and 3.2.3. The analysis follows the same structure as the analysis in both Sections 4.3 and 5.3, i.e., we use the same tools and figures. For a more detailed overview as well as introduction into tables and figures, the reader is referred to the simulation study in Chapter 4.

In this section we first consider an ex post analysis in Section 6.2.1. In the ex post analysis, we choose the one best combination of hyperparameters for each test set ex post. In Section 6.2.2 we consider the results of the out-of-sample analysis. In the out-of-sample analysis, we estimate the best hyperparameter combination for each observation in the test set by cross-validation. To this end, we use a rolling window of size 40 and the last ten observations are used as a validation set.

Candidate Values for Hyperparameters For forecast combination with individual feature bounds there are many hyperparameters we need to define or estimate. We can choose the transformation function Ψ, the features (or rather their function) we evaluate forecasts on $\boldsymbol{\nu}$, and the smallest and largest feature deviation (ψ_{min}, ψ_{max}). We can choose those separately for the individual lower and upper bounds. Additionally, there are between zero and three additional arguments that are required depending on the transformation function. As a result, the number of candidate tuples of hyperparameters can get very large fast and, thus, it is very computational demanding.

Therefore, we limited the number of candidate hyperparameter values to evaluate as follows:

- smallest deviation: $\psi_{min} \in \{0, 0.1, 0.2, 0.3\}$, see Section 6.1.1.
- largest deviation: $\psi_{max} \in \{\psi_{min} + 0.1, \psi_{min} + 0.2, \psi_{min} + 0.3\}$, see Section 6.1.1.
- transformation function: $\Psi \in \{step, GRelU\}$, see Sections 6.1.2.2 and 6.1.2.3.
- feature: $\xi() \in \{MSE, AvgMSEC, AccDiv\}$, see Equations (2.2), (6.46) and (6.47).
- feature values: $\boldsymbol{\nu} \in \{ranked, not\ ranked\}$, see Section 6.1.1.

It is important to note that we have to limit the hyperparameter values that we consider. For ψ_{min} and ψ_{max}, we believe that the added value of very large weights is limited and concentrate on smaller weights that are closer to the prior weights. Evidence that supports this, is the good forecasting performance of the PW method. By design, it limits each weight to be at most equal to one. It has noticeably better forecasting performance than OW, see for example the result in Tables 4.1 and 4.2. We decided to define the largest deviation ψ_{max} relative to the lower deviation as it prevents infeasible constellations where $\psi_{max} < \psi_{min}$ most effectively.

For a more detailed discussion on why we chose the *step* and *GRelU*, see again the concluding remarks in Section 6.1.2.4. With respect to the features, we use measures for the accuracy, diversity as well as the combination of them that we proposed. Thereby, we can analyze the value each feature provides. Lastly, we use both ranked and the actual feature values. By that, we can have constraints that are driven by the raw feature values which can be distributed far from equally. In contrast, for ranked feature values the feature are distributed equally and, therefore, the individual feature bounds are distributed along the range of result provided by the transformation function.

We evaluate the hyperparameter values listed above individually for both the lower and upper individual feature bound. As a result, for both the ex post and out-of-sample analysis we evaluate 1728 unique setups of hyperparameters.

6.2.1 Ex Post Analysis: Individual Feature Bounds

In this section we discuss the result of forecast combination with individual feature bounds with respect to all scenarios in Section 6.2.1.1. Thereafter, we consider the results with respect to the error correlations, error variance similarity, and special groups in Section 6.2.1.2.

Forecast combination with individual feature constraint is related to bounded weights as we discussed in Section 6.1. Both methods constraint weights by imposing lower and upper bounds, however, in case of Individual Feature Bounds the bounds are individual for each weight. Because they are related, we will also consider the following results of Individual Feature Bounds in relation to Bounded Weights.

6.2.1.1 Ex Post Analysis: Overall Results

Tables 6.1 and 6.2 present the results for each scenario of the simulation study. To this end, we calculated the average MSE over all test sets scenario-wise. It is designed similar to Tables 4.1, 4.2, 5.3 and 5.4.

CM	z	SG	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
1	0.05	none	0.98	0.97	2.11	0.98	**0.86**	**0.86**	**0.86**	**0.86**
		first	0.97	0.93	1.89	0.96	**0.81**	**0.81**	**0.81**	**0.81**
		last	1.01	0.99	2.04	1.00	**0.85**	**0.85**	**0.85**	**0.85**
		both	0.97	0.93	1.75	0.96	**0.77**	**0.77**	**0.77**	**0.77**
1	0.20	none	1.16	0.92	1.03	1.12	0.50	**0.49**	**0.49**	0.50
		first	1.12	0.73	0.52	1.04	0.25	**0.24**	**0.24**	**0.24**
		last	1.24	0.95	0.84	1.18	0.40	**0.39**	**0.39**	**0.39**
		both	1.15	0.72	0.45	1.05	0.22	**0.21**	**0.21**	**0.21**
1	0.50	none	1.53	0.80	0.42	1.33	0.20	**0.19**	**0.19**	0.20
		first	1.38	0.30	0.09	0.85	**0.04**	**0.04**	**0.04**	**0.04**
		last	1.64	0.82	0.35	1.38	0.17	**0.16**	**0.16**	**0.16**
		both	1.46	0.30	0.08	0.85	**0.04**	**0.04**	**0.04**	**0.04**
2	0.05	none	0.56	0.62	1.37	0.56	**0.53**	0.55	0.55	**0.53**
		first	0.56	0.62	1.34	0.56	**0.52**	0.54	0.54	**0.52**
		last	0.57	0.64	1.37	0.57	**0.54**	0.56	0.55	**0.54**
		both	0.57	0.63	1.35	0.57	**0.53**	0.55	0.55	**0.53**
2	0.20	none	0.68	0.68	1.42	0.66	**0.58**	**0.58**	**0.58**	**0.58**
		first	0.64	0.56	1.11	0.60	**0.47**	0.48	**0.47**	0.48
		last	0.68	0.66	1.31	0.65	0.55	**0.54**	**0.54**	0.55
		both	0.66	0.54	1.00	0.60	**0.44**	**0.44**	**0.44**	**0.44**
2	0.50	none	0.87	0.64	1.10	0.76	0.50	**0.49**	**0.49**	0.50
		first	0.80	0.26	0.35	0.49	0.18	0.18	0.18	**0.17**
		last	0.96	0.66	1.05	0.81	**0.48**	**0.48**	**0.48**	**0.48**
		both	0.83	0.26	0.33	0.49	**0.17**	**0.17**	**0.17**	**0.17**
3	0.05	none	**0.25**	0.32	0.63	0.26	0.26	0.27	0.27	**0.25**
		first	0.25	0.31	0.63	0.25	0.25	0.26	0.26	**0.24**
		last	**0.25**	0.32	0.63	**0.25**	0.26	0.27	0.27	**0.25**
		both	0.25	0.32	0.64	0.25	0.25	0.27	0.26	**0.24**
3	0.20	none	0.30	0.36	0.71	0.30	**0.29**	0.30	0.30	**0.29**
		first	0.29	0.32	0.63	0.27	**0.26**	0.27	**0.26**	**0.26**
		last	0.30	0.36	0.71	0.29	**0.28**	0.29	0.29	**0.28**
		both	0.30	0.32	0.61	0.27	**0.25**	0.26	0.26	0.26
3	0.50	none	0.40	0.39	0.75	0.35	**0.31**	0.32	0.32	0.32
		first	0.35	0.19	0.35	0.22	**0.16**	0.17	0.18	0.17
		last	0.43	0.40	0.76	0.36	**0.32**	0.33	**0.32**	0.33
		both	0.37	0.20	0.36	0.22	**0.16**	0.17	0.18	0.17

Table 6.1. Simulation study results of benchmark and Individual Feature Weights methods forecast combination methods for correlation matrices CM1, CM2 and CM3 (ex post analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

CM	z	SG	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
4	0.05	none	0.63	0.69	1.57	0.63	**0.60**	0.63	0.63	0.61
		first	0.65	0.70	1.61	0.65	**0.61**	0.64	0.64	0.62
		last	0.66	0.72	1.64	0.66	**0.62**	0.65	0.65	0.63
		both	0.64	0.68	1.59	0.63	**0.60**	0.62	0.62	**0.60**
4	0.20	none	0.79	0.78	1.78	0.76	**0.70**	0.71	**0.70**	**0.70**
		first	0.73	0.67	1.48	0.68	**0.60**	**0.60**	**0.60**	**0.60**
		last	0.80	0.78	1.74	0.77	**0.70**	**0.70**	**0.70**	**0.70**
		both	0.77	0.68	1.51	0.70	**0.61**	**0.61**	**0.61**	**0.61**
4	0.50	none	1.03	0.79	1.50	0.91	**0.68**	**0.68**	**0.68**	**0.68**
		first	0.90	0.30	0.45	0.56	**0.26**	**0.26**	**0.26**	**0.26**
		last	1.10	0.79	1.47	0.93	0.69	**0.68**	**0.68**	0.70
		both	1.02	0.31	0.46	0.60	0.27	**0.26**	0.27	0.27
5	0.05	none	0.44	0.40	0.84	0.45	**0.33**	0.36	0.36	0.34
		first	0.44	0.40	0.85	0.45	**0.33**	0.36	0.36	0.35
		last	0.44	0.41	0.88	0.46	**0.34**	0.37	0.36	0.35
		both	0.43	0.41	0.87	0.45	**0.34**	0.36	0.36	0.35
5	0.20	none	0.51	0.51	1.06	0.54	**0.43**	0.45	0.45	**0.43**
		first	0.48	0.49	0.98	0.50	0.42	0.43	0.42	**0.41**
		last	0.52	0.54	1.13	0.56	0.47	0.48	0.48	**0.46**
		both	0.49	0.53	1.06	0.51	0.45	0.46	0.45	**0.44**
5	0.50	none	0.67	0.66	1.21	0.69	0.57	0.57	0.57	**0.55**
		first	0.56	0.31	0.42	0.45	0.27	**0.26**	0.27	**0.26**
		last	0.66	0.67	1.22	0.68	0.58	0.58	0.58	**0.56**
		both	0.59	0.33	0.44	0.46	0.29	0.28	0.28	**0.27**
6	0.05	none	0.79	0.63	1.39	0.77	**0.54**	0.56	0.56	0.55
		first	0.78	0.60	1.32	0.75	**0.51**	0.53	0.53	0.52
		last	0.80	0.63	1.37	0.78	**0.54**	0.56	0.55	0.55
		both	0.80	0.61	1.30	0.77	**0.52**	0.53	0.53	0.53
6	0.20	none	0.97	0.64	1.16	0.89	**0.49**	**0.49**	0.50	**0.49**
		first	0.94	0.52	0.95	0.81	0.41	**0.40**	**0.40**	0.41
		last	1.02	0.64	0.96	0.91	**0.43**	**0.43**	**0.43**	**0.43**
		both	0.97	0.51	0.80	0.81	0.36	**0.35**	**0.35**	**0.35**
6	0.50	none	1.28	0.61	0.85	1.02	0.38	**0.37**	0.38	0.38
		first	1.22	0.26	0.42	0.62	0.19	**0.18**	**0.18**	**0.18**
		last	1.39	0.62	0.69	1.05	0.31	**0.30**	**0.30**	0.31
		both	1.31	0.26	0.38	0.62	0.17	**0.16**	**0.16**	0.17

Table 6.2. Simulation study results of benchmark and Individual Feature Weights methods forecast combination methods for correlation matrices CM4, CM5 and CM6 (ex post analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

The result immediately show that forecast combination with individual feature constraints improves the forecast accuracy in comparison to the benchmark methods.

The percentages of how often each method has, potentially among others, the smallest MSE, the average rank and distances are shown in Table 6.3 similar to Tables 4.3

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
Smallest MSE (%)	2.78	0.00	0.00	1.39	61.11	47.22	47.22	**62.50**
Avg Rank	6.58	5.71	7.34	5.98	2.37	2.94	2.87	**2.22**
Avg Distance	0.35	0.14	0.58	0.25	**0.00**	0.01	0.01	**0.00**

Table 6.3. Key figures for the MSE values of benchmark and Individual Feature Bounds methods over all simulation study scenarios (ex post analysis). Smallest MSE (%) - Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank - Average rank of a method where a smaller rank is favorable. Avg Distance - Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

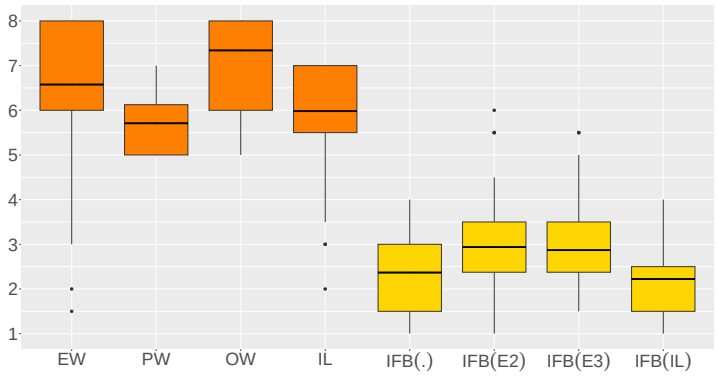


Figure 6.9. Boxplot of ranks across benchmarks and Individual Feature Bounds methods for the ex post analysis.

and 5.5.[48] Overall, both IFB(·) and IFB($\dot{\boldsymbol{\omega}}^{IL}$) have the smallest MSE for a majority of scenarios (61.11% and 62.50% respectively). Using the other shrinkage directions $\dot{\boldsymbol{\omega}}^{EW2}$ and $\dot{\boldsymbol{\omega}}^{EW3}$ has the smallest MSE for 47.22% of scenarios. They are oftentimes but not always the best method for the same scenarios.

With respect to the ranks of the methods, IFB(·) and IFB($\dot{\boldsymbol{\omega}}^{IL}$) have the smallest average ranks. Their general superiority becomes evident when considering Figure 6.9. It shows a boxplot of ranks across all scenarios.

Although, the methods have similar average and also median ranks (black horizontal bar in the boxplot), the middle 50% of IFB($\dot{\boldsymbol{\omega}}^{IL}$) have a noticeably smaller range of values. Accordingly, it has lower ranks more regularly. For 75% of scenarios, its rank is at most 2.5. Beside the small average rank of IFB and IFB($\dot{\boldsymbol{\omega}}^{\boldsymbol{IL}}$) they have an average distance of zero to the best method of zero. Although, this is somewhat expected as they

[48] A comparison of average ranks with the results in Chapters 4 and 5 is not sensible as they depend on the number of methods considered.

are the best method oftentimes, it nevertheless shows their consistently good forecast performance.[49]

Overall all Individual Feature Bounds methods have a superior forecast accuracy compared to the benchmarks. IFB(·) and in particular IFB($\dot{\boldsymbol{\omega}}^{\boldsymbol{IL}}$) are the favorable approaches.

If we consider the Individual Feature Bounds methods as a whole, i.e., both IFB(·) and IFB($\dot{\boldsymbol{\omega}}$), for all but two scenarios ($CM3/0.05/none$ and $CM3/0.05/last$) at least one IFB method has the *strictly* smallest MSE value.

6.2.1.2 Ex Post Analysis: Groups of Scenarios

After evaluating the overall results of the forecast combination methods in the previous section, we now analyze the results with respect to error correlation matrices, error variance similarities and special groups. Table 6.4 shows how often each method has, potentially among others, the smallest MSE for each correlation matrix.

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
CM1	0.00	0.00	0.00	0.00	50.00	**100.00**	**100.00**	83.33
CM2	0.00	0.00	0.00	0.00	**75.00**	50.00	58.33	**75.00**
CM3	16.67	0.00	0.00	8.33	**66.67**	0.00	16.67	58.33
CM4	0.00	0.00	0.00	0.00	**83.33**	58.33	58.33	58.33
CM5	0.00	0.00	0.00	0.00	41.67	8.33	0.00	**66.67**
CM6	0.00	0.00	0.00	0.00	50.00	**66.67**	50.00	33.33

Table 6.4. Percentage of scenarios for which benchmarks and Individual Feature Bounds methods have the smallest MSE with respect to the error correlation matrix (ex post). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

Overall, only EW and IL have the smallest MSE for some scenarios (CM3 - small error correlations 0.2). The results show that the two best methods overall (IFB(·) and IFB($\dot{\boldsymbol{\omega}}^{IL}$)) do not have a particular error correlation for which they are not suited for. However, they are the best methods least often for CM5 and CM6 respectively. Recall that in CM5 the better a group of forecast is, the higher its error correlation is and for CM6 the best forecasts have a medium error correlation (0.5) while all others have highly correlated forecast errors (0.9). With respect to their average ranks, they are quite consistent across all error correlation matrices. The only exception is IFB($\dot{\boldsymbol{\omega}}^{IL}$). It has a significantly smaller rank for CM5, see Figure 6.10.

[49] For comparison, in the out-of-sample analysis in Section 5.3.2, the PW approach was similarly often the best method as IFB($\dot{\boldsymbol{\omega}}^{EW2}$) from the ex post analysis of this chapter. However, the average distance of PW was 0.07 while here IFB($\dot{\boldsymbol{\omega}}^{EW2}$) has an average distance of 0.01, see Tables 5.10 and 6.3.

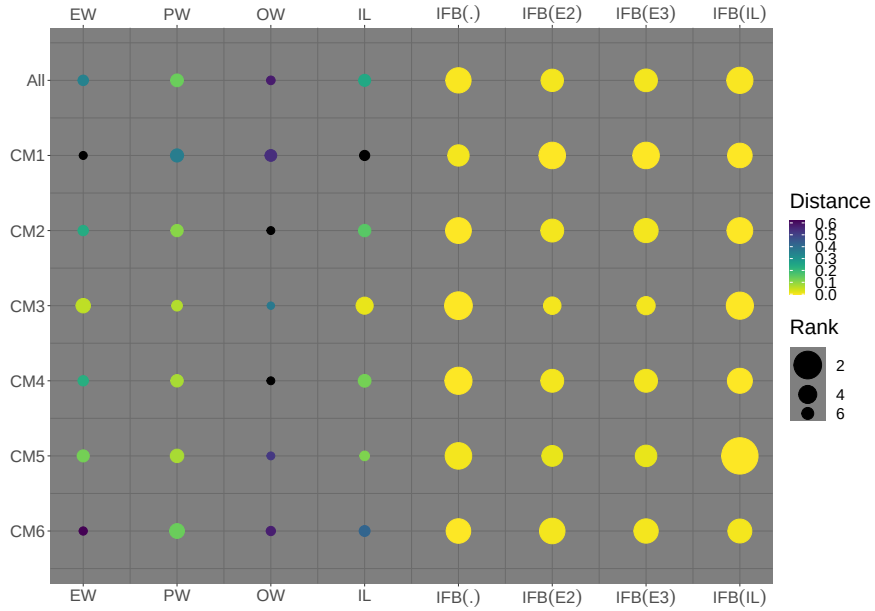


Figure 6.10. Illustration of average ranks and distances of the benchmarks and Individual Feature Bounds methods for different correlations matrices (ex post analysis).

Interestingly, the shrinkage directions we proposed to use $\boldsymbol{\omega}^{EW2}$ and $\boldsymbol{\omega}^{EW3}$ (Conditional Group Equal Weights) have quite different results compared to $\boldsymbol{\omega}^{IL}$. For comparison, for Bounded Weights (BW) the result of the prior weights are more similar, see Table 5.6 and Figure 5.8. The difference is most noticeably both in Figure 6.10. IFB($\boldsymbol{\omega}^{EW2}$) and IFB($\boldsymbol{\omega}^{EW3}$) have the smallest MSE for none, one or two scenarios for CM3 and CM5. In contrast, IFB($\boldsymbol{\omega}^{IL}$) is the best method for more than half of these scenarios and has noticeably better ranks. However, similar to the result from Bounded Weights, IFB($\boldsymbol{\omega}^{EW2}$) and IFB($\boldsymbol{\omega}^{EW3}$) are noticeably the best method for highly correlated forecasts errors (CM1) and even mostly favorable for CM6.

Table 6.5 shows how often each method has the smallest MSE with respect to the error variance similarities and special groups (rows). To complement the analysis Figure 5.9 depicts the average ranks and distance to the best method scenario-wise. For IFB(·) we see similar results as for BW(·) when considering the error variance similarity. For both IFB(·) and BW(·) we see that they have less often the smallest MSE as forecasts have more diverse or dissimilar error variance, i.e., forecasting performances. Shrinkage towards prior weights again is more often superior as forecasts become more diverse with respect to their error variance, see Tables 5.7 and 6.5. However, again IFB($\boldsymbol{\omega}^{IL}$) does not follow the same pattern. For both $z = 0.05$ and $z = 0.50$ it has, among other methods, the smallest MSE for about half of scenarios. However, for $z = 0.20$ is has the smallest MSE for 19 out of 24 scenarios.

Comparing again IFB(·) and BW(·), BW(·) is more often the best method with respect to the error variance similarity compared to IFB(·), the opposite is true if prior

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
$z = 0.05$	8.33	0.00	0.00	4.17	**83.33**	16.67	16.67	54.17
$z = 0.20$	0.00	0.00	0.00	0.00	58.33	58.33	66.67	**79.17**
$z = 0.50$	0.00	0.00	0.00	0.00	41.67	**66.67**	58.33	54.17
none	5.56	0.00	0.00	0.00	**66.67**	44.44	38.89	55.56
first	0.00	0.00	0.00	0.00	61.11	44.44	50.00	**66.67**
last	5.56	0.00	0.00	5.56	55.56	50.00	55.56	**61.11**
both	0.00	0.00	0.00	0.00	61.11	50.00	44.44	**66.67**

Table 6.5. Percentage of scenarios for which the benchmarks and Individual Feature Bounds methods have the smallest MSE with respect to the error variance similarity and special groups (ex post analysis). The total number of scenarios for each error variance similarity is 24 and for special groups it is 18. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

weights are used. They are the best method more often if individual feature bounds around prior weights are used instead of common bounds.

With respect to the average ranks depicted in Figure 6.11, a similar pattern is visible, i.e., IFB(·) has noticeably better ranks if the error variances are more similar while the opposite holds if prior weights are used. Except for IFB($\dot{\boldsymbol{\omega}}^{IL}$) which is noticeably more consistent than other methods (better rank) for all error variances similarities. With respect to special groups there is again no particular pattern visible. This becomes apparent if Figure 6.11 is considered. The average ranks do not change noticeably with respect to special groups and in comparison to the average ranks over *all* scenarios that are depicted on the top of the figure.

6.2.2 Out-Of-Sample: Individual Feature Bounds with Hyperparameter Estimation

In this section we analyze the out-of-sample forecast accuracy of the forecast combination methods proposed in this chapter, i.e., individual feature bounds. To this end, we determine the hyperparameter a priori based on past information. Additionally, we apply cross-validation for each observation in the test, i.e., we re-estimate which combination of hyperparameters is best repeatedly, see again Section 2.1. Recall that this can lead to a smaller MSE compared to the ex post analysis where we chose a single combination of hyperparameters for all observations in the test set.

In what follows we first analyze the MSE result overall scenarios in Section 6.2.2.1. In Section 6.2.2.2 we analyze the forecast accuracy with respect to the error correlations, error variance similarity and special groups.

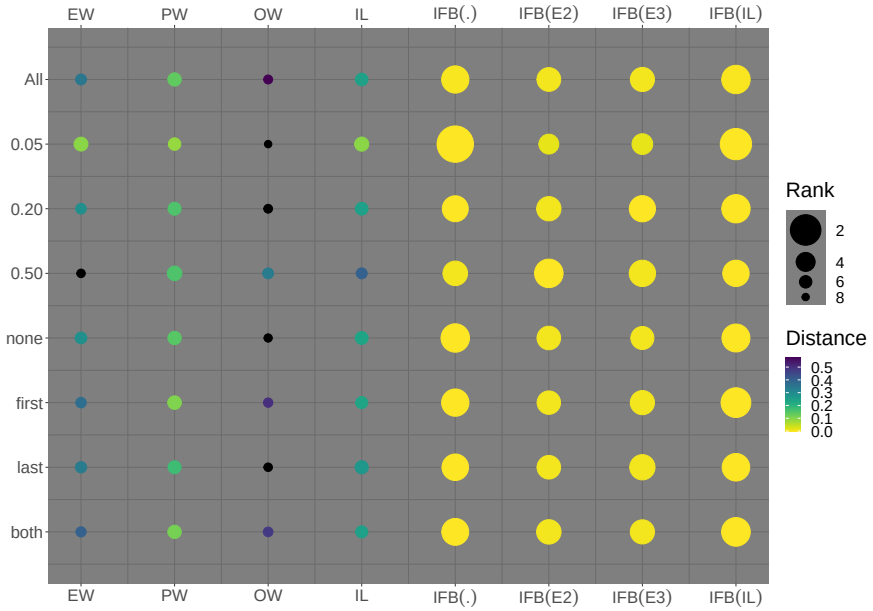


Figure 6.11. Illustration of average ranks and distances of the benchmarks and Individual Feature Bounds methods for error variance similarities and special groups (ex post analysis).

6.2.2.1 Out-Of-Sample: Overall Results

The MSE result for each scenario are in Tables 6.7 and 6.8. At first glance, similarly to the result from BW, the superiority of the sophisticated forecast combination methods in comparison to the benchmarks is not apparent anymore.

This observation is supported by the methods percentage of having the smallest MSE, average and range of ranks as well as distance present in both Table 6.6 and Figure 6.12.

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
Smallest MSE (%)	23.61	**38.89**	0.00	29.17	8.33	25.00	22.22	12.50
Avg Rank	4.89	**2.68**	7.34	4.03	5.10	4.10	4.07	3.79
Avg Distance	0.27	0.07	0.50	0.18	0.07	0.06	0.06	**0.05**

Table 6.6. Key figures for the MSE values of benchmark and Individual Feature Bounds methods over all simulation study scenarios (out-of-sample analysis). Smallest MSE (%) - Percentage of scenarios for which the method has the smallest MSE, potentially among others. Avg Rank - Average rank of a method where a smaller rank is favorable. Avg Distance - Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

CM	z	SG	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
1	0.05	none	0.98	**0.97**	2.11	0.98	1.02	1.01	1.01	1.01
		first	0.97	**0.93**	1.89	0.96	0.95	0.94	0.94	0.95
		last	1.01	**0.99**	2.04	1.00	1.00	1.00	1.00	1.00
		both	0.97	0.93	1.75	0.96	**0.92**	**0.92**	**0.92**	**0.92**
1	0.20	none	1.16	0.92	1.03	1.12	0.64	**0.63**	**0.63**	**0.63**
		first	1.12	0.73	0.52	1.04	0.33	**0.32**	**0.32**	0.33
		last	1.24	0.95	0.84	1.18	0.52	**0.51**	**0.51**	0.52
		both	1.15	0.72	0.45	1.05	0.30	**0.29**	**0.29**	**0.29**
1	0.50	none	1.53	0.80	0.42	1.33	0.28	**0.27**	**0.27**	0.28
		first	1.38	0.30	0.09	0.85	0.07	0.07	0.07	**0.06**
		last	1.64	0.82	0.35	1.38	0.24	**0.23**	**0.23**	**0.23**
		both	1.46	0.30	0.08	0.85	**0.06**	**0.06**	**0.06**	**0.06**
2	0.05	none	**0.56**	0.62	1.37	**0.56**	0.67	0.69	0.68	0.67
		first	**0.56**	0.62	1.34	**0.56**	0.68	0.69	0.68	0.67
		last	**0.57**	0.64	1.37	**0.57**	0.71	0.71	0.70	0.69
		both	**0.57**	0.63	1.35	**0.57**	0.69	0.69	0.69	0.68
2	0.20	none	0.68	0.68	1.42	**0.66**	0.78	0.75	0.75	0.76
		first	0.64	**0.56**	1.11	0.60	0.66	0.63	0.63	0.64
		last	0.68	0.66	1.31	**0.65**	0.74	0.71	0.71	0.72
		both	0.66	**0.54**	1.00	0.60	0.62	0.58	0.59	0.60
2	0.50	none	0.87	**0.64**	1.10	0.76	0.70	0.67	0.67	0.69
		first	0.80	0.26	0.35	0.49	**0.25**	0.26	0.27	0.26
		last	0.96	0.66	1.05	0.81	0.69	**0.65**	**0.65**	0.67
		both	0.83	0.26	0.33	0.49	**0.25**	**0.25**	0.26	**0.25**
3	0.05	none	**0.25**	0.32	0.63	0.26	0.35	0.35	0.34	0.32
		first	**0.25**	0.31	0.63	**0.25**	0.35	0.34	0.33	0.32
		last	**0.25**	0.32	0.63	**0.25**	0.35	0.35	0.34	0.33
		both	**0.25**	0.32	0.64	**0.25**	0.35	0.34	0.33	0.32
3	0.20	none	**0.30**	0.36	0.71	**0.30**	0.41	0.40	0.38	0.38
		first	0.29	0.32	0.63	**0.27**	0.38	0.35	0.35	0.35
		last	0.30	0.36	0.71	**0.29**	0.40	0.38	0.37	0.37
		both	0.30	0.32	0.61	**0.27**	0.37	0.34	0.33	0.34
3	0.50	none	0.40	0.39	0.75	**0.35**	0.46	0.43	0.42	0.43
		first	0.35	**0.19**	0.35	0.22	0.24	0.24	0.24	0.23
		last	0.43	0.40	0.76	**0.36**	0.47	0.44	0.43	0.44
		both	0.37	**0.20**	0.36	0.22	0.24	0.24	0.25	0.23

Table 6.7. Simulation study results of benchmark and Individual Feature Bounds method forecast combination methods for correlation matrices CM1, CM2 and CM3 (out-of-sample analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

CM	z	SG	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
								$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
4	0.05	none	**0.63**	0.69	1.57	**0.63**	0.75	0.75	0.74	0.73
		first	**0.65**	0.70	1.61	**0.65**	0.76	0.75	0.75	0.74
		last	**0.66**	0.72	1.64	**0.66**	0.78	0.78	0.78	0.76
		both	0.64	0.68	1.59	**0.63**	0.75	0.74	0.73	0.72
4	0.20	none	0.79	0.78	1.78	**0.76**	0.89	0.85	0.85	0.84
		first	0.73	**0.67**	1.48	0.68	0.77	0.72	0.72	0.73
		last	0.80	0.78	1.74	**0.77**	0.88	0.84	0.84	0.84
		both	0.77	**0.68**	1.51	0.70	0.80	0.74	0.74	0.75
4	0.50	none	1.03	**0.79**	1.50	0.91	0.90	0.85	0.85	0.86
		first	0.90	**0.30**	0.45	0.56	0.35	0.34	0.35	0.35
		last	1.10	**0.79**	1.47	0.93	0.91	0.86	0.86	0.88
		both	1.02	**0.31**	0.46	0.60	0.37	0.36	0.37	0.37
5	0.05	none	0.44	**0.40**	0.84	0.45	0.47	0.48	0.48	0.46
		first	0.44	**0.40**	0.85	0.45	0.46	0.48	0.48	0.46
		last	0.44	**0.41**	0.88	0.46	0.47	0.49	0.49	0.46
		both	0.43	**0.41**	0.87	0.45	0.47	0.49	0.49	0.46
5	0.20	none	**0.51**	**0.51**	1.06	0.54	0.60	0.60	0.60	0.58
		first	**0.48**	0.49	0.98	0.50	0.59	0.57	0.56	0.55
		last	**0.52**	0.54	1.13	0.56	0.64	0.64	0.64	0.61
		both	**0.49**	0.53	1.06	0.51	0.61	0.60	0.60	0.57
5	0.50	none	0.67	**0.66**	1.21	0.69	0.81	0.76	0.76	0.73
		first	0.56	**0.31**	0.42	0.45	0.39	0.36	0.37	0.35
		last	**0.66**	0.67	1.22	0.68	0.79	0.76	0.75	0.73
		both	0.59	**0.33**	0.44	0.46	0.42	0.38	0.39	0.37
6	0.05	none	0.79	**0.63**	1.39	0.77	0.68	0.70	0.70	0.69
		first	0.78	**0.60**	1.32	0.75	0.66	0.67	0.67	0.66
		last	0.80	**0.63**	1.37	0.78	0.69	0.69	0.70	0.69
		both	0.80	**0.61**	1.30	0.77	0.67	0.67	0.68	0.67
6	0.20	none	0.97	0.64	1.16	0.89	0.64	**0.63**	0.64	0.64
		first	0.94	**0.52**	0.95	0.81	0.53	**0.52**	**0.52**	0.53
		last	1.02	0.64	0.96	0.91	0.57	**0.55**	**0.55**	0.56
		both	0.97	0.51	0.80	0.81	0.48	**0.47**	**0.47**	0.48
6	0.50	none	1.28	0.61	0.85	1.02	0.51	**0.50**	**0.50**	0.51
		first	1.22	**0.26**	0.42	0.62	**0.26**	**0.26**	**0.26**	**0.26**
		last	1.39	0.62	0.69	1.05	0.44	**0.43**	**0.43**	0.44
		both	1.31	0.26	0.38	0.62	**0.24**	**0.24**	**0.24**	**0.24**

Table 6.8. Simulation study results of benchmark and Individual Feature Bounds method forecast combination methods for correlation matrices CM4, CM5 and CM6 (out-of-sample analysis). The table depicts the MSE of the forecast combination method. The methods with the smallest MSE are depicted in bold numbers.

Using the weights determined by PW for forecast combination result in the smallest MSE most often (38.89%). Both EW (23.61%) and IL (29.17%) also have a better forecast accuracy more often than any of the Individual Feature Bounds methods except for IFB($\dot{\boldsymbol{\omega}}^{EW2}$) (25%). Although using the IFB($\dot{\boldsymbol{\omega}}$ with shrinkage directions $\dot{\boldsymbol{\omega}}^{EW2}$ and

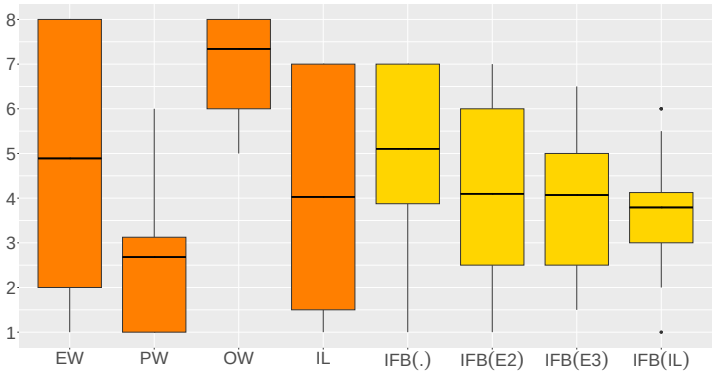


Figure 6.12. Boxplot of ranks across the benchmarks and Individual Feature Bounds methods for the pseudo out-of-sample analysis.

$\dot{\boldsymbol{\omega}}^{EW3}$ achieve lower ranks for some scenarios, see Figure 6.12, $\dot{\boldsymbol{\omega}}^{IL}$ is more consistent, i.e., the range of its middle 50% of ranks is smaller. However, PW has both the smallest rank and most favorable range of ranks as one can see by the middle 50% of ranks. With respect to the distance to the best methods scenario-wise, the Individual Feature Bounds methods are however closer to it than PW, except for IFB(·), i.e., without prior weights. Overall, the result are again similar to what we observed for BW, see Table 5.10. Note that we compare the percentages of being the best methods and the distance and not the average ranks. The latter is incomparable by definition, because the ranks are relative to the methods MSE and their overall number.

6.2.2.2 Out-Of-Sample: Groups of Scenarios

With respect to the error correlation matrices using Conditional Group Equal Weights, i.e., IFB($\dot{\omega}$), is more often superior for error correlations matrices that include highly correlated forecasts, see Table 6.9. They have the smallest MSE for about two thirds of scenarios. IFB($\dot{\boldsymbol{\omega}}^{IL}$) is favorable for half of scenarios for CM1 and only a sixth of scenarios for CM6. Shrinkage towards prior weights is particularly useful for highly correlated error matrices. This finding is similar to what we observed for bounded weights. However, IFB($\dot{\omega}$) methods have the smallest MSE values more often compared to BW($\dot{\omega}$), see Table 5.11.

The average ranks in Figure 6.13 provide further evidence that the shrinkage directions $\dot{\boldsymbol{\omega}}^{EW2}$ and $\dot{\boldsymbol{\omega}}^{EW3}$ are well-suited for highly correlated forecast errors. For all other correlations matrices, i.e., medium (0.5 CM2), small (0.2 CM3) and mixed error correlations, benchmark methods provide better MSE values more often and have better average ranks (larger circle) and distances (yellow color), see Figure 6.13.

With respect to the error variances similarity, IFB(·) has the smallest MSE for $z = 0.05$ only for one scenario. Similar to BW(·) the more diverse or dissimilar forecasts are in terms of their error variances, the more often IFB(·) is the better method. The

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
CM1	0.00	25.00	0.00	0.00	16.67	**66.67**	**66.67**	50.00
CM2	33.33	25.00	0.00	**50.00**	16.67	16.67	8.33	8.33
CM3	41.67	16.67	0.00	**75.00**	0.00	0.00	0.00	0.00
CM4	25.00	**50.00**	0.00	**50.00**	0.00	0.00	0.00	0.00
CM5	41.67	**66.67**	0.00	0.00	0.00	0.00	0.00	0.00
CM6	0.00	50.00	0.00	0.00	16.67	**66.67**	58.33	16.67

Table 6.9. Percentage of scenarios for which benchmarks and Individual Feature Bounds methods have the smallest MSE with respect to the error correlation matrix (out-of-sample analysis). The total number of scenarios for each correlation matrix is twelve. The methods with the highest percentages for each correlation matrix are depicted in bold numbers.

	EW	PW	OW	IL	IFB(·)	IFB($\dot{\omega}$)		
						$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
$z = 0.05$	**45.83**	**45.83**	0.00	**45.83**	4.17	4.17	4.17	4.17
$z = 0.20$	20.83	25.00	0.00	**33.33**	0.00	**33.33**	29.17	8.33
$z = 0.50$	4.17	**45.83**	0.00	8.33	20.83	37.50	33.33	25.00
none	27.78	**38.89**	0.00	33.33	0.00	22.22	16.67	5.56
first	22.22	**55.56**	0.00	22.22	11.11	16.67	16.67	11.11
last	27.78	22.22	0.00	**38.89**	0.00	27.78	27.78	5.56
both	16.67	**38.89**	0.00	22.22	22.22	33.33	27.78	27.78

Table 6.10. Percentage of scenarios for which the benchmarks and Individual Feature Bounds methods have the smallest MSE with respect to the error variance similarity and special groups (out-of-sample analysis). The total number of scenarios for each error variance similarity is 24 and for special groups it is 18. The methods with the highest percentages for each error variance similarity are depicted in bold numbers.

average ranks depicted in Figure 6.14 lead to a similar conclusion. However, using individual feature bounds around prior weights overall has a better average rank and distance compared to using the IFB(·).

The comparison with the results from BW($\dot{\omega}$) shows that the uncertainty introduced by hyperparameter estimation is more sever for IFB($\dot{\omega}$). BW($\dot{\omega}$) is the best method more often than IFB($\dot{\omega}$) if both are compared with the benchmarks, see Table 5.11. For example, for $z = 0.5$ the best BW($\dot{\omega}$) methods has the smallest MSE for 41.67% of scenarios. In comparison, for the best IFB($\dot{\omega}$) method the proportion is 37.50. If we compare IFB(·) and BW(·), however, using individual feature constraint is better more often ($z = 0.5$). With respect to special groups, we do not see a clear pattern.

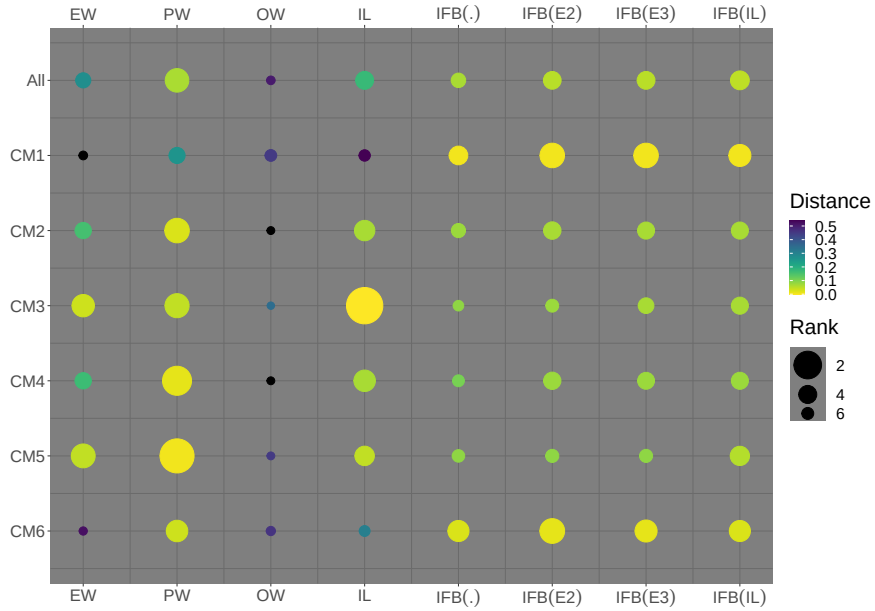


Figure 6.13. Illustration of average ranks and distances of the benchmarks and Individual Feature Bounds for different correlations matrices (out-of-sample analysis).

6.3 Discussion and Future Work

In this thesis, we analyzed forecast combination methods with constrained weights, i.e., we shrink weights towards a particular direction, including prior weights. We considered L_1 methods that shrinks weights by imposing one constraint for all weights in Chapter 4. Due to the fact that the L_1 constraint generally allows for single forecasts or weights to have a larger effect on the combined forecast than others, we introduced Bounded Weights in Chapter 5 that diversifies the combined forecast but allows for individual weights to improve the forecast accuracy. To this end, we use both a common lower and common upper bounds in general or around prior weights.

In this chapter, we consider both research questions stated in Chapter 1: *how to further improve the forecast accuracy of a combined forecast using constrained weights* and *how to incorporate additional, external information in forecast combination with constrained weights?* To this end, introduced forecast combination with Individual Feature Bounds which is the fifth main contribution of this thesis. Individual Feature Bounds are related to Bounded Weights as we also impose lower and upper bounds. However, each forecast or weight gets individual lower and upper bounds that are determined by a feature value or characteristic of themselves. The more favorable the feature value of a forecast is, the more lose its constraint is. In other words, forecast with unfavorable feature values get constraint to be closer to the shrinkage direction or prior weights.

In Section 6.1 we introduced the idea of Individual Feature Bounds and defined the optimization problem. This includes, first and foremost, the Individual Feature De-

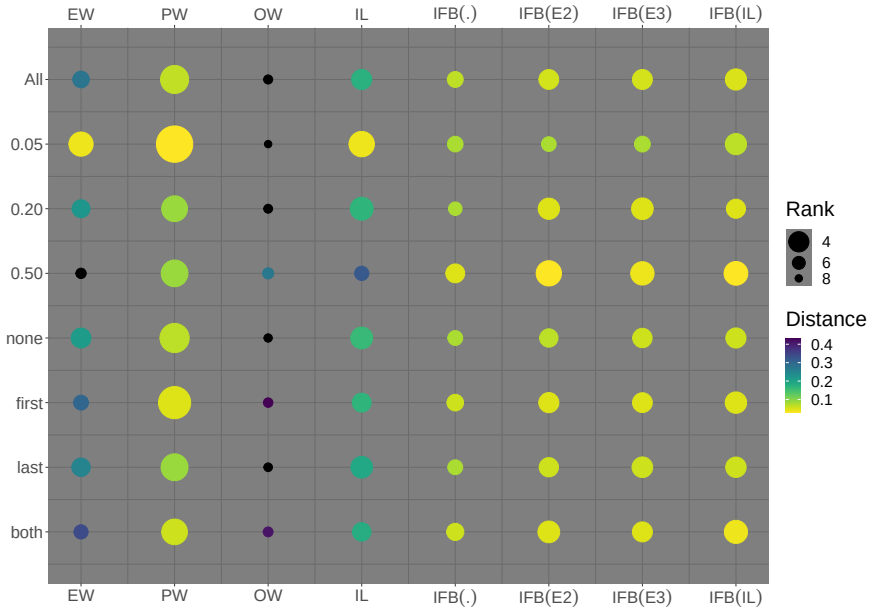


Figure 6.14. Illustration of average ranks and distances of the benchmarks and Individual Feature Bounds methods for error variance similarities and special groups (out-of-sample analysis).

viation that, together with the prior weights, defines the Individual Feature Bounds (Objective I of this chapter).

In Section 6.1.1 we discuss the components that are needed to define the Individual Feature Deviation: the transformation function and the feature values. Among other parameters, for the transformation function we can define the smallest and largest deviation that they have as an output. The feature values are scaled to ensure a standardized input into the transformation function. To this end, we either use ranks or the raw feature values. As a result of the former Individual Feature Deviations are distributed equally across the possible outcomes of the transformation function. If the raw feature values are used, it depends on the actual distribution of the feature values how weights are constraint.

Section 6.1.2 present four transformation function: linear, step, GRelU and the generalized logistics function. These transformation functions have properties that result in different Individual Feature Deviations and, by that, Bounds. For example, the step function can be used to select certain weights to be equal to their shrinkage direction or prior weight which also allows for a variable selection. The GRelU imposes a common bound for a certain amount of weights that have less favorable feature values. All other forecasts are linearly less constraint the more favorable their feature value is.

In Section 6.1.3 we introduce feature values that are based on two concepts: forecast accuracy and diversity. The former, measure the performance of a forecast. The latter, how unique or diverse it is compared to the other forecasts. For forecast accuracy we

used the MSE and for diversity the MSEC. We also proposed use the AvgMSEC between forecasts in order make it possible to consider the MSEC as a measure of diversity for Individual Feature Bounds. Additionally, we proposed the measure AccDiv that combines both accuracy and diversity into one feature.

With respect to both the ex post analysis in Section 6.2.1 and the out-of-sample analysis in Section 6.2.2 we found in general similar result to Bounded Weights. However, the use of Individual Feature Bounds improves the forecast accuracy compared to Bounded Weights for both the ex post and out-of-sample analysis. In particular the result from the out-of-sample analysis are interesting. Individual Feature Bounds have more hyperparameters, and we used a curated set of values for them. Nevertheless, although due to the larger number of hyperparameters Individual Feature Bounds could be more strongly influenced by the estimation uncertainty of hyperparameters, our new method improves the forecast accuracy.

Although, we compare the benchmarks, LHS, L_1, Bounded Weights and Individual Feature Bounds methods in Chapter 7 based on real-world data, we want to very briefly compare their forecast accuracies in the simulation study. Table 6.11 shows how often each method as a group has, among other methods, the smallest MSE across all scenarios. The first rows show the result of the ex post and the second row the result of the out-of-sample analysis. For the ex post analysis, the Individual Feature Bounds is with-

	Bench.	LHS	L_1	BW	IFB
Ex-post	2.78	2.78	9.72	9.72	**97.22**
Out-of-sample	**68.06**	11.11	18.06	12.50	13.89

Table 6.11. Percentage of scenarios for which the benchmark, LHS, L_1, Bounded Weights and Individual Feature Bounds methods have the smallest MSE over all scenarios.

out questions the best approach. It has the smallest MSE in almost all scenarios. Both L_1 and Bounded Weights are better than the benchmarks. For the out-of-sample analysis, as expected, the benchmarks are oftentimes the best methods. Bounded Weights and Individual Feature Bounds less often have the smallest MSE compared to L_1.

This again shows a major problem of the forecast combination methods with hyperparameter estimation. Although, they can have a superior forecast accuracy, the uncertainty introduced by the estimation of hyperparameters prohibits the methods to use their full potential. Nevertheless, overall the shrinkage methods or forecast combination with constrained weights are better than the benchmark even with hyperparameter estimation for about a third of scenarios. For future research we suggest again to analyze hyperparameter estimation to utilizes the potential of the considered methods, in particular the Individual Feature Bounds. Additionally, we suggest that another extensive simulation study should be conduced that compares the benchmarks, LHS,

L_1, Bounded Weights and Individual Feature Bounds together. Due to the already extensive analysis of the simulation studies this is beyond the scope of this thesis.

Instead, we will focus on an extensive empirical analysis comparing the out-of-sample forecasting performance off all considered methods based on real-world data in Chapter 7.

With respect to the second research question of this thesis: *how to incorporate additional, external information in forecast combination with constrained weights* we used Individual Feature Bounds. However, one can argue that prior weights themselves also introduce external information into the forecast combination problem. For example, we could use a similar formula as the inverse-loss weighted average of Equation (4.16) but with different features to derive prior weights. Additionally, instead of using Individual Feature Bounds another way could be to deviate from the original objective function and create new objective functions that incorporates the additional information directly into the objective function. For example, consider again the example that we want to create a combined forecast for food retail that is particularly tailored towards promotional periods. We could design an optimization problem where the objective function is extended by a term that measure the forecast accuracy particularly for those periods. We do believe that our proposed method of Forecast Combination with Individual Feature Bounds can incorporate various external information more conveniently. Moreover, the transformation function allow us to enforce properties such as variable selection. Additionally, the focus of this thesis is forecast combination with constrained weights in form of the original forecast combination problem, i.e., quadratic objective function with linear constraints. However, for future research, we strongly suggest analyzing the usefulness of prior weights based on feature values and customized objective functions to incorporate additional information.

In summary, Individual Feature Bounds is a new direction for forecast combination. While existing approaches estimate weights based on feature values, we define constraints based on them. This follows the concept of the original forecast combination method intended by J. M. Bates and Granger (1969). The weights are estimated such that they minimize the error variance of the combined forecast, however, subject to additional constraint. The method we propose has great potential. It can be used in any forecast context or application and, most importantly, incorporate application-specific, external information. We strongly suggest that future research about Individual Feature Bounds should analyze its forecasting performance for multiple forecast application with and without application-specific features.

7 Empirical Analysis

In this chapter we analyze the forecasting performance of the so far presented forecast combination methods on real world data. Our goal is to ensure a fair and objective assessment of the methods. Therefore, we do not select only a few time series, but evaluate the methods for numerous time series. For comparison, in the related literature around the L_1 constraint Diebold and Shin (2019) and Radchenko et al. (2023) use one and Roccazzella et al. (2022) use two time series or datasets to evaluate the forecast combination methods.

With respect to the overall structure of this thesis, this chapter completes the last main contribution stated in Chapter 1: extensive simulation studies and a comprehensive empirical analysis. In particular the empirical analysis enables us to evaluate and assess the usefulness our considered and newly proposed forecast combination methods in real-word scenarios. Because we are using numerous time series, we ensure that our results are meaningful and more robust.

In our empirical analysis we use the first 1000 monthly time series of the M4 dataset (Makridakis et al., 2018, 2020).[50] The M4 Competition featured times series of different frequencies (yearly, quarterly, monthly et cetera) and areas (economics, industry, finance et cetera). The objective of the competition was to evaluate how forecasting methods and algorithms perform on real-world data in a competitive scenario and learn from the results. To this end, for each time series participants had to provide forecasts for the whole test set at once. The test set was unknown to them.

We chose the M4 dataset because it is well-known, recognized and publicly available for anyone. However, we do not intend to follow the same procedure as the M4 competition which required participants to provide forecast based on the training data for a certain forecast horizon at once (18 for monthly series). Within this thesis we focused on one-step ahead forecasts and, therefore, we will also use this forecast horizon for the observations of the test set. Accordingly, we do not intend to compare our the result of our analysis to the M4 competition results.

In Section 3.2 we analyzed the considered and proposed forecast combination methods separately using our simulation study. In this chapter, we consider the forecast accuracy of all methods at once. Thereby, we analyze if one is generally superior or if all methods provide value but for certain time series.

[50] Downloaded from `https://github.com/Mcompetitions/M4-methods` (06.03.2023).

Figure 7.1. Exemplary time series from the first 1000 monthly M4 dataset.

The remainder of this chapter is organized as follows. In Section 7.1 we briefly present the data and the necessary forecasting process. This includes the creation of input forecasts and how we combine those forecasts. In Section 7.2, we analyze the result of the forecast combination methods. This includes, the forecast accuracy over all time series in Section 7.2.1, a brief analysis of the hyperparameter values in Section 7.2.2 and the forecast accuracy with respect to certain characteristics in Section 7.2.3. Lastly, we summarize and discuss our results in Section 7.3.

7.1 Data and Procedure

For our analysis we chose to use the first 1000 monthly time series of the M4 dataset. The time series are of different length and scale. For each time series there is a test set that consists of 18 month of observations, i.e., 1.5 years. In the following descriptive analysis we exclude the test set observations.

On average the first 1000 monthly times series of the M4 data set have 324 observations. The shortest time series consists of 61 months (about 5 years), while the longest spans over 1230 month, i.e., 102.5 years. The middle 50% of time series have a length between 209 and 330, i.e., about roughly 17 to 28 years of monthly observations.

Figure 7.1 depicts five of the time series exemplary. Because the time series are of different length, the abscissa shows a time index that is one for the first observation of a time series. Accordingly, the time index of the last observation corresponds to the length of the time series. The ordinate shows the values of the time series. Time series M893 depicted by the pink-colored line has a clearly visible seasonal pattern as well as a trend. In contrast, M679 (green) has no clear trend but a seasonality. However,

the seasonality is less regular compared to M893 and the oscillations or fluctuations get larger over time, i.e., the variance of the time series increases over time. M679 also appears to have some periodic behavior, see time index one to roughly 120 then until 240.[51] A more clear periodic behavior is present for M201 (olive). The periods are roughly $5-6$ years long. However, M201 time series has a less regular course or path in comparison to M893 and M679. Moreover, for one of the periods (time index 120 to 180) it has a noticeable increase. M120 (red) is a longer time series with about 69 years of monthly observations. Note that it has a long term trend and a structure break or significant drop at roughly 730. This could for example be a major event like the financial crisis. Lastly, M843 (blue) is, in contrast, one of the shorter time series that spans about 67 months, i.e., about 5.5 years. Due to the different scales of the time series it may appear that it is relatively constant, however it fluctuates and has a clear trend.

These examples show that the monthly M4 dataset consists of a variety of time series with different properties. This is very valuable for the evaluation of forecast combination methods. We can measure and compare the performance of the methods to each other for a diverse set of time series instead of special cases. Moreover, this has the potential to identify which method is well-suited for which kind of time series.

Forecasting Methods In order to use the first 1000 monthly time series of the M4 data set for forecast combination, we first need forecasts that then can be combined. To this end, we follow Montero-Manso et al. (2020) who proposed FFORMA. Recall, that FFORMA is a meta-learning approach that uses characteristics of time series to determine weights of forecasts. It placed second in the M4 competition (Makridakis et al., 2018, 2020). In this thesis we use the same forecasting methods and functions to generate the input forecasts for our forecast combination methods:[52]

1. *Naive* — The forecast is equal to the last observation, i.e., $\hat{y}_{t+1} = y_t$ (Petropoulos et al., 2022).

2. *Random Walk with Drift* — The forecast is equal to the last observation and a drift term, i.e., $\hat{y}_{t+1} = y_t + \mu$ (Petropoulos et al., 2022).

3. *Seasonal Naive* — The forecast is equal to the last observation of the same season as the target period, i.e., $\hat{y}_{t+1} = y_{t-o+1}$ where o is the length of a season (Petropoulos et al., 2022).

[51] Note that we use the term seasonal pattern here in context of the monthly series, i.e., a season corresponds to a year. If a time series has a longer systematic course of rising / falling values, we use the terms period or periodic.

[52] It has to be noted that these forecasting methods are well-known and commonly used methods. The reason for the placement of the FFORMA approach is not likely to be just because of the used input methods but due to the sophisticated algorithm for creating weights that combine the forecasts. Accordingly, we think that using these forecasts with the knowledge that they can be combined well for multiple forecast horizons does not distort our analysis.

4. *Theta Method* — A method that decomposes the time series in different parts, modifies them separately using a theta coefficient, extrapolates and then combines them (Assimakopoulos & Nikolopoulos, 2000).

5. *ARIMA* — A model that consists of an autoregressive and moving average part. It can use differences of the time series to ensure stationarity. In can be extended to a *SARIMA* model which additionally includes seasonal components (Hyndman & Khandakar, 2008; Petropoulos et al., 2022).

6. *Exponential Smoothing* — At its core (simple) exponential smoothing is a weighted average of past observations where the weight decreases exponential the older the observation is. In its more complex form it includes a trend and a (additive or multiplicative) seasonal component (Petropoulos et al., 2022).

7. *TBATS model* — An advanced forecasting method based on exponential smoothing that uses i.a. Box-Cox transformations, Fourier representations as well as ARMA error corrections. It is designed to forecast time series that have complex seasonal patterns (de Livera, Hyndman, & Snyder, 2011).

8. *STLM-AR* — A procedure that applies a time series decomposition (STL) to deseasonalize the data, models, and forecasts it via an autoregressive model, i.e., only the autoregressive part of an ARMA, and then re-seasonalizes the forecast (Cleveland, Cleveland, McRae, & Terpenning, 1990; Hyndman et al., 2023).

9. *Neural Network* — A feed forward neural network with a single hidden layer that uses lagged values of the time series as input variables. A neural network is a system of weighted sums and biases that are transformed by an activation function (Petropoulos et al., 2022).

We generate all forecasts using functions from the *R* package *"forecast"* (Hyndman et al., 2023) with the default settings, see also Hyndman and Khandakar (2008); Montero-Manso et al. (2020).

Forecast and Forecast Combination Process The objective of this chapter is to generate out-of-sample (combined) forecasts for the M4 time series and evaluate the forecasting performance. To this end, we need to develop a process that covers how to

1. train forecast models and generate forecasts,

2. train the forecast combination methods based on the generated forecasts,

3. estimate hyperparameter based on a validation set and

4. compute forecasts for the test set.

In the following process, we apply pseudo out-of-sample forecasting, estimate hyperparameters via cross-validation, i.e., a validation set, and use both the expanding window and rolling window approach as described in Section 2.1.

The process we designed is illustrated in Figure 7.2. The illustration shows how we

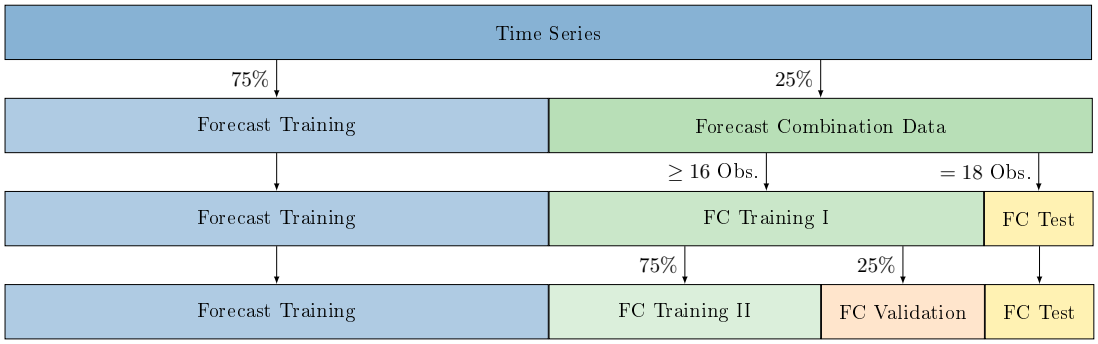


Figure 7.2. Illustration of the process to create forecasts and perform forecast combination (FC) and partitioning of each of the 1000 monthly time series from the M4 data set used. The size of the rectangles do not represent proportions and are for illustration purposes only.

segment time series into different parts. The uppermost rectangle depicts the whole time series, including the observations from the test set.

First, we need to train forecasting models and create forecasts. To this end, we split the data into *Forecast Training* (blue) and *Forecast Combination Data* (green), see Figure 7.2. The former consists of the first 75% of observations, and is used to train the nine forecast methods presented above.[53] Then, we use those methods to generate forecasts for the remaining 25% of observations from the time series with an expanding window, recall Section 2.1 and Figure 2.1. The resulting forecasts from this are the data that can be used for the forecast combination. It is depicted by the green rectangle in the second row of Figure 7.2. Again, the objective of the analysis is to evaluate and compare the performance of the forecast combination methods. Therefore, we decided to use a 75/25 split because we want to ensure that the input forecasts have higher quality, i.e., a better forecast accuracy. To this end, there needs to be enough training data to train the forecast methods. For the same reason we decided to use an expanding window for generating the input forecasts. Thereby, we always use as much data as available to train forecast models and generate forecasts. Additionally, as a result of the 75/25 split, we can analyze something that we assess to be of greater value. Within the simulation study we did not vary the amount of training data for the forecast combination method. However, as we discussed, the hyperparameter estimation of the proposed forecast combination methods prevents the methods from reaching their best potential forecast accuracy. Accordingly, now we can also analyze how an

[53] If the number of observations is not an integer, we round down to the nearest integer.

increasing training set for the forecast combination methods effects their out-of-sample performance, however, across different time series.

As a result of the first step of our process, we have forecasts for about the last 25% of the time series. This is depicted by the green rectangle called *Forecast Combination Data* in Figure 7.2. It includes the test set which consists of 18 observations. Accordingly, we have to split *Forecast Combination Data*, into a training and test set. The former is depicted by the green rectangle in the third row of Figure 7.2 called *FC Training I*. The latter is depicted by the yellow rectangle called *FC Test* and includes the last 18 observation of each time series. Although, we intended a 75/25 split between *Forecast Training* and *Forecast Combination Data* it is, however, not always feasible because some time series are too short. The shortest monthly time series has 79 month of observations, which includes the test set observations. 25% of that correspond to about 20 observation from which 18 are the test set. Based on that, weights can not be estimated. To ensure that weights for nine input forecasts can be estimated and there are spare observation to estimate hyperparameter using cross-validation, we reduce the amount of training data for the forecasting method, if necessary, such that *Forecast Combination Data* includes at least 34 observations. 16 observations for weight estimation and the hyperparameter search (*FC Training I*) plus the original 18 observations of the test set (*FC Test*), see again Figure 7.2.

To estimate the hyperparameters for the forecast combination methods we use cross-validation and a rolling window, see again Chapter 2. To this end, we split *FC Training I* into *FC Training II* (green) and *FC Validation* (orange) also using a 75/25 split (Figure 7.2). Accordingly, for our analysis we have time series where hyperparameters estimation is based on as few as 4 observations while for other time series we use about 73 observations, i.e., over six years of monthly observations. Because this discrepancy in available data is one focus of our analysis, we use a rolling window to forecast the observations of the test set with the forecast combination methods. Thereby, the number of available observation for weight and hyperparameter estimation is constant for the test set of a given time series.

In summary, we train forecasting models on about 75% of the time series (*Forecast Training*) and generate forecasting for the remaining 25% using an expanding window. We create combined forecasts for the 18 observations in the test set using a rolling window. For each observation in the test set we estimate hyperparameters based on a validation set which includes 25% of the forecast combination training set.

7.2 Results

In this section we analyze the results of forecasting methods for the M4 time series.[54]

Prior to that, we want to emphasize that it is possible that some of the nine forecasting methods have the same forecasts. If this happens for a whole training set, it can lead to perfect multicollinearity and, as a result, the calculation of the inverse error variance covariance matrix as it is implemented is not possible. The exact same forecast adds no value to the forecast combination methods. Therefore, we remove one of them. It was the case for 26 time series. Moreover, for about eleven time series one or multiple forecast combination methods reported an error. For the sake of a fair comparison between forecast combination methods, we removed those eleven time series from the results, i.e., we analyze the result of 989 monthly time series.

To compare the forecast accuracy of the forecast combination methods across different time series, we can not use the MSE. The time series are of different scale or magnitude and the MSE is a scale-dependent measure, recall Section 6.1.3.1. Instead, we will use the relative MSE (relMSE). To this end, for each time series we divide the MSE of each forecast combination methods by the MSE of EW. Accordingly, a relMSE of less (greater) than one indicates that the corresponding method has a smaller (larger) MSE compared to equal weights.

In what follows, we consider the overall performance of the methods for all time series in Section 7.2.1. Thereafter, Section 7.2.2 briefly analyzes the estimated hyperparameters. Lastly, in Section 7.2.3 we analyze the results with respect to forecast features. Similar to the analysis of the simulation study, we consider the error correlation of forecasts and the similarity of the empirical error variance. Because these empirical time series are more complex and of larger variety, we have to use alternative key figures, i.e., the mean error correlation of forecasts and the variation in in-sample forecast accuracy (MSE). We do not consider special groups, because the impact on the forecast combination methods have been limited in the simulation studies from Chapters 4 to 6. In addition to the mean error correlation and variation in MSE, we also consider the length of the time series as it influences both the estimation uncertainty of weights and hyperparameters.

7.2.1 Overall Forecast Accuracy

Figure 7.3 illustrates the relMSE values (ordinate) over all 989 time series for each method (abscissa) as a boxplot. Recall, the box represents the middle 50%, i.e., the values between the 25%- and 75%-Quantile. The black horizontal bar depicts the median. The horizontal line throughout Figure 7.3 shows the relMSE of the EW, which

[54] Due to the large amount of time series and methods, we summarize the result and do not depict the relMSE for each individual time series.

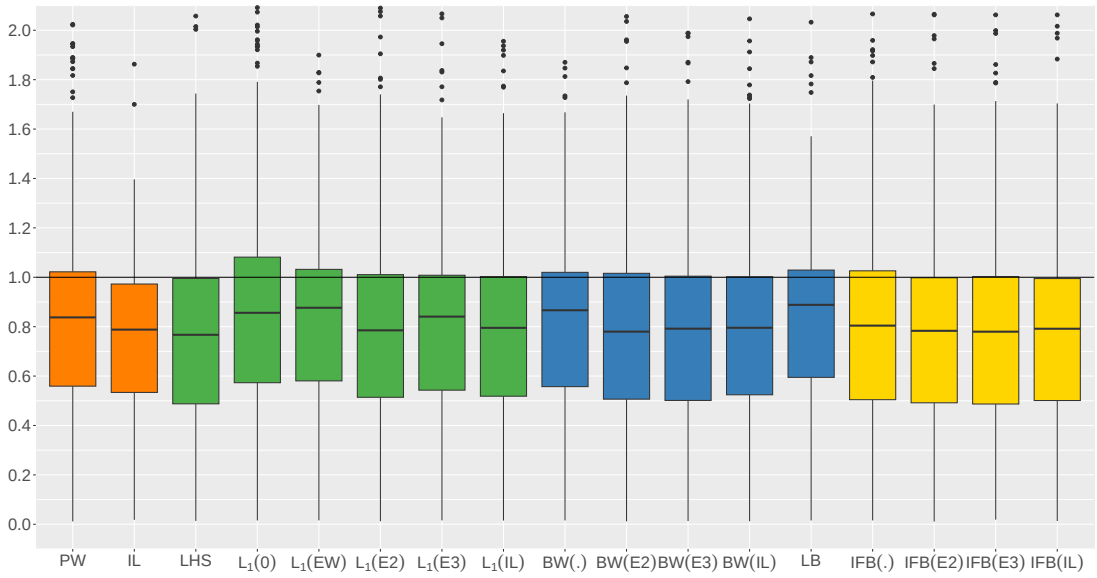


Figure 7.3. Boxplot of the relMSE values of the M4 time series for all considered forecast combination methods. Difference in relative MSE to the best method for each M4 time series. The ordinate has been limited and OW and LS have been removed for a better visibility.

is one by definition. The forecast combination methods are grouped and differentiated by color. The benchmark methods are orange, the LHS and L_1 methods green, blue for the BW methods and yellow for the IFB methods. Note that we omitted both OW and LS from Figure 7.3 because they have very large relMSE values and those distort the figure and the analysis, e.g., the 75%-Quantile is about 98.[55] However, due to the fact that there are very large relMSE values present in the results, we will henceforth use the median instead of the mean if the relMSE is considered. The median is more robust to outliers or very large values and, thus, the analysis is not distorted as it would be using the mean.

Figure 7.3 shows that there is no method that achieves significantly smaller relMSE values regularly and over all time series in comparison to the other methods. Shrinkage towards zero using an L_1 constraint, i.e., $L_1(0)$, has a noticeably higher 75%-Quantile, and it is greater one. Accordingly, the EW is better suited combining forecasts for certain time series. For PW, $L_1(1/N)$, $L_1(\boldsymbol{\omega}^{E2})$, $L_1(\boldsymbol{\omega}^{E3})$, BW$(\cdot)$, BW$(\boldsymbol{\omega}^{E2})$, LB and IFB$(\cdot)$ the 75%-Quantile is slightly above one. All other approaches have a smaller or equal relative MSE than the EW for roughly 75% of times series. IL is the only method where the 75%-Quantile is clearly smaller one and its 25%-Quantile together with IFB$(\boldsymbol{\omega}^{E2})$ and IFB$(\boldsymbol{\omega}^{E3})$ is the smallest among all methods.

[55] Recall, that LS nests OW, and we estimate the shrinkage parameter by the formula provided by Blanc and Setzer (2020). This results in almost the same forecasts as OW.

	PW	OW	IL	LHS		$L_1(\kappa)$		$L_1(\dot{\boldsymbol{\omega}})$		
				LS	LHS	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
Median relMSE	0.84	6.00	0.79	6.00	0.77	0.86	0.88	0.79	0.84	0.80
	BW(·)	BW($\dot{\omega}$)			LB	IFB(·)	IFB($\dot{\omega}$)			
		$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$			$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$	
Median relMSE	0.87	0.78	0.79	0.80	0.89	0.80	0.78	0.78	0.79	

Table 7.1. Median relMSE of the considered monthly M4 time series for all forecast combination methods.

In addition to Figure 7.3, Table 7.1 present the values of the median relMSE (rows) for each method (columns). With respect to the median relMSE, the LHS method has the smallest value (0.77) followed by IFB($\boldsymbol{\omega}^{E2}$), IFB($\boldsymbol{\omega}^{E3}$) and BW($\boldsymbol{\omega}^{E2}$) with 0.78. In general, shrinkage towards prior weights (except $L_1(\boldsymbol{\omega}^{E3})$) and IFB(·) have small median relMSE values between 0.78 and 0.8. Shrinkage towards a fixed value $L_1(\kappa)$ as well as BW(·) and LB have noticeably larger median relMSE values (0.86 to 0.89). In contrast to the out-of-sample analysis from Section 5.3.1, BW(·) is better than LB although two hyperparameters estimated. The PW approach that was very competitive in the out-of-sample analysis of L_1, BW and IFB methods, is somewhere in-between the methods with a median relMSE (0.84).

Table 7.2 presents how often each method has one of the smallest relMSE values, the average rank and median distance to the best method for a given time series. We still use the average ranks, because ranks are not effected by large relMSE value. In addition to Table 7.2, Figures 7.4 and 7.5 show boxplots of the ranks and distance for each method. Again, we removed OW and LS from Figure 7.5. The three methods that have the smallest relMSE value most often are: LHS, IL and IFB(·) with 21.03%, 12.54% and 10.82% of the time series. As we have roughly 1000 time series, multiplying these number with ten gives roughly the actual number of time series for which these methods have, potentially among other, the smallest relMSE.

In particular LHS turns out to be a favorable method for a variety of time series. It also has the second-smallest rank and among the smallest median distance to the best method. Similar to the result from the simulation study, the inverse-loss weighted average is a competitive forecast combination method. It is the preferable methods more often than equal weights. This becomes even more apparent looking at Figures 7.4 and 7.5. The distribution of ranks and distances of EW is noticeably larger (higher box) compared to the other methods. PW and IFB($\dot{\omega}$) with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ have the smallest relMSE for roughly 10% of the time series. OW and LS have the smallest percentage of being among the best methods with 2.33%.

For shrinkage towards prior weights the results show that the shrinkage direction we proposed to use within a forecast combination optimization problem, i.e., $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$,

Method		Best (%)	Avg Rank	Med Dist.
EW		8.09	13.76	0.31
PW		10.01	9.41	0.09
OW		2.33	18.48	5.30
IL		12.54	8.35	0.07
LHS	LS	2.33	17.67	5.30
	LHS	**21.03**	8.12	**0.06**
$L_1(\kappa)$	0	4.95	11.34	0.12
	$1/N$	4.65	11.09	0.12
$L_1(\dot{\boldsymbol{\omega}})$	$\boldsymbol{\omega}^{E2}$	7.38	9.07	0.08
	$\boldsymbol{\omega}^{E3}$	6.07	9.77	0.09
	$\boldsymbol{\omega}^{IL}$	4.25	9.46	0.08
BW($\cdot$)		4.35	10.70	0.11
BW($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	8.80	9.00	0.08
	$\boldsymbol{\omega}^{E3}$	7.99	8.98	0.08
	$\boldsymbol{\omega}^{IL}$	6.27	9.46	0.08
LB		3.94	11.46	0.12
IFB($\cdot$)		10.82	9.21	0.07
IFB($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	9.50	8.25	0.07
	$\boldsymbol{\omega}^{E3}$	9.61	**8.11**	**0.06**
	$\boldsymbol{\omega}^{IL}$	7.68	8.30	0.07

Table 7.2. Key figures of each forecast combination method for the 989 monthly M4 time series. Smallest MSE — Percentage of time series for which the method has the smallest MSE, potentially among others. Avg Rank — Average rank of a method where a smaller rank is favorable. Avg Distance — Average distance or difference in MSE the method and best method scenario-wise. The method with the most favorable value are depicted in bold numbers.

for all $L_1(\dot{\boldsymbol{\omega}})$, BW($\dot{\boldsymbol{\omega}}$) and IFB($\dot{\boldsymbol{\omega}}$) more often have a smaller relMSE than shrinking weights towards $\boldsymbol{\omega}^{IL}$, see Table 7.2. However, with respect to Figures 7.4 and 7.5 there is less of a difference between shrinkage directions for IFB($\dot{\boldsymbol{\omega}}$) (similar distribution of ranks and distance). Moreover, for $L_1(\dot{\boldsymbol{\omega}})$ shrinkage towards $\boldsymbol{\omega}^{E3}$ has a larger IQR compared to $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{IL}$ and a higher 75%-Quartile, i.e., it has higher ranks and is further from the best method more often. BW($\cdot$), LB and shrinkage towards a fixed value $L_1(\kappa)$ are noticeably less often the best method, have higher average ranks and median distance than BW($\dot{\omega}$), IFB($\dot{\omega}$) and IFB($\cdot$), see Figures 7.4 and 7.5.

With respect to the forecast combination methods we proposed, IFB($\dot{\omega}$) (and to an extent IFB($\cdot$)) are the best. They are the best method, potentially among other, most often, have (among) the smallest average ranks and median distances. IFB($\dot{\omega}$) with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$ has the third and smallest average rank across all methods. In conjunction with the fact that IFB($\dot{\omega}$) also has the smallest values for the median distance (together with LHS) provides evidence, that it is a more reliable and consistent method compared to others. This is supported by Figures 7.4 and 7.5. The middle 50% include smaller ranks and their 75%-Quantiles are smaller than other methods, except for the 75%-Quantile of BW($\boldsymbol{\omega}^{E2}$) which is about identical. For the median distance it is less apparent that

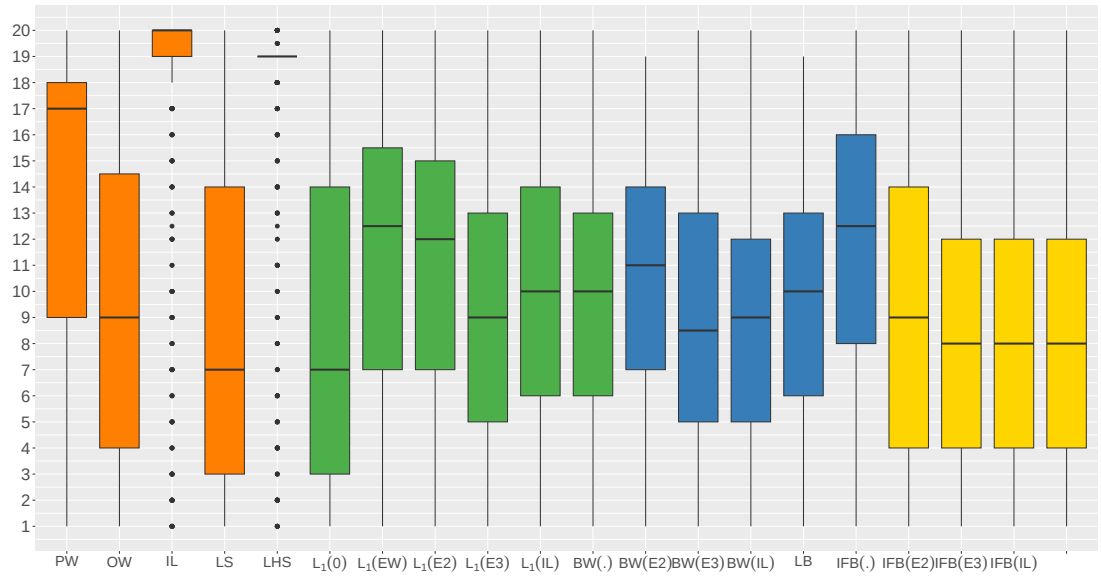


Figure 7.4. Ranks of the relative MSE values for each M4 time series and all considered forecast combination methods. The ordinate has been limited for a better visibility.

the middle 50% are closer to zero. However, one can see that the 75%-Quantiles are the smallest among the proposed method together with IL (about 0.125). Although, LHS can have smaller ranks and distances, the distribution of the middle 50% is wider. Accordingly, for some time series it also has noticeably larger ranks and distances.

We would like to emphasize that, again, we are comparing all methods against each other. This leads to a distorted perception, as one underestimates the group of shrinkage and the methods we proposed. If we compare the benchmark methods (EW, PW, OW, and IL) with the shrinkage methods considered in this thesis, it turns out that the shrinkage methods (all of LHS, L_1, BW and IFB methods) as a group are superior for 67% of the time series. Note that for each time series, we compare the smallest relMSE of the benchmark with the smallest relMSE of the shrinkage methods. Let us now consider only the approaches we proposed in this thesis, i.e.,$L_1(\dot{\omega})$ with $\boldsymbol{\omega}^{E2}$ and $\boldsymbol{\omega}^{E3}$, BW$(\cdot)$, BW$(\dot{\omega})$, IFB$(\cdot)$ and IFB$(\dot{\omega})$. They are superior to the benchmark methods for 63% of the time series. For these time series, our methods median improvement in relMSE is 7.3% (18% average).

We can also consider the methods benchmarks, LHS, L_1, BW and IFB methods as whole. Table 7.3 shows how often each method has the smallest relMSE (%) and strictly smallest MSE (%).[56] Both the benchmark methods and IFB have, among other

[56]Note that percentages for smallest MSE add up to values greater 100 as there can be multiple methods with the same relMSE value since we rounded relMSE values to the second decimal place. Furthermore, strictly smallest MSE adds up to less than 100, because we consider how often each method has the strictly smallest relMSE. If two methods from different groups have the same relMSE, the corresponding time series counts for neither of them.

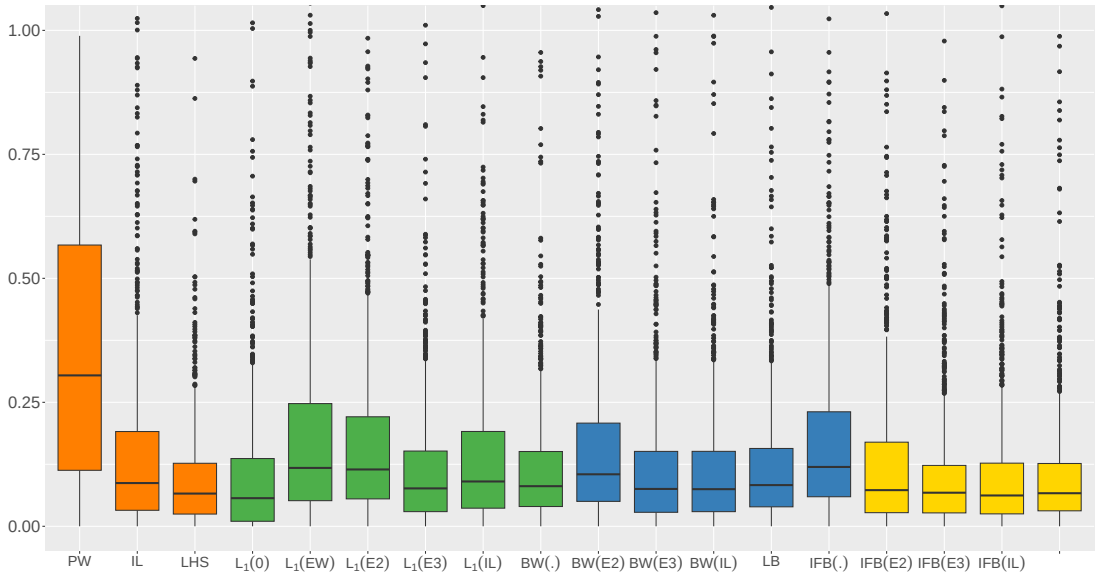


Figure 7.5. Difference in relative MSE to the best method for each M4 time series. We limited the ordinate and removed OW and LS for a better visibility.

	Benchmark	LHS	L_1	BW	IFB
Smallest MSE (%)	31.65	23.15	22.55	24.67	**32.05**
Strictly smallest MSE (%)	20.02	13.04	9.61	10.41	**22.24**

Table 7.3. Percentage of time series for which the benchmarks, LHS, L_1, Bounded Weights and Individual Feature Bounds methods have the smallest and strictly smallest MSE.

methods, the smallest MSE for roughly 32% of time series, i.e., about 317 out of 989. For the other methods this holds for about 23% (L_1) to 25% (BW) of time series.

Forecast combination with Individual Feature Bounds have the strictly smallest relMSE most often (22.24%). L_1 and BW have the strictly smallest MSE for about 10% of time series, LHS is strictly the best for 13%, and he benchmarks for about 20% of time series.

This evidence demonstrates clearly that the newly proposed forecast combination methods complement existing approaches, in particular our proposed new field of forecast combination where we constrain forecast individually based on feature values (Individual Feature Bounds).

This process can potentially be enhanced with both information about the estimated hyperparameter values and characteristics of the corresponding datasets.

7.2.2 Hyperparameter Breakdown

In this section we take a look at the hyperparameters of the considered forecast combination methods. To this end, we consider all hyperparameters over all test set observations for each method. Then we compute the median value over all hyperparameter values.

Again, we use the median value as it robust to outliers or very large values. By using the mean, a few very large observations can distort the analysis. The median enables a more fair comparison. For example, for $L_1(0)$ the mean value over the test sets estimated value for γ_0 is 4.21. However, this corresponds roughly to the 83%-quantile of estimated hyperparameter values. For 83% of test set observations the hyperparameter value were smaller than the average. In contrast, the median value over all test sets hyperparameters is 1.1, i.e., for half of the time series, the median hyperparameter was smaller or equal to 1.1.[57]

LHS methods The hyperparameter of LS confirm what we expect based on the relMSE of this method, weights are not substantially shrunken towards EW. Instead, the hyperparameter is so small that it basically result in the same weights as OW. In contrast, for LHS the median hyperparameter is one (0.95 average). Accordingly, the best 50% of weights are close to their corresponding equal weights while the other are close to zero. Nonetheless, small deviation from those weights are present for some time series, if it turns out to improve forecasts accuracy.

L_1 methods Table 7.4 shows the median shrinkage parameter and *coefficient of variation (CoefVar)* over all test set observations. The coefficient of variation is the scaled, dimensionless standard deviation. Thereby, we can compare the variation in estimated hyperparameters that are defined on different scales or intervals, e.g., $L_1(0)$ and $L_1(1/N)$. The closer the coefficient of variation is to zero, the smaller the variation in estimated hyperparameters. To this end, one divides the standard deviation by the mean value, see Fahrmeir et al. (2016, p. 68).

	$L_1(\kappa)$		$L_1(\dot{\omega})$		
	0	$1/N$	$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
Median	1.10	0.60	0.20	0.20	0.20
CoefVar	2.10	2.52	2.99	3.01	2.93

Table 7.4. L_1 methods: Median and coefficient of variation (CoefVar) of the shrinkage parameter γ for the considered M4 time series.

Recall, that the smallest feasible value for $L_1(0)$, i.e., γ_0, is one, while for all other methods it is zero. Overall, the shrinkage parameter is close to the smallest feasible values. For example, with $\gamma_0 = 1.1$ the largest possible weight is 1.05. Otherwise,

[57] It needs to be noted that we use estimated hyperparameters of test set observations for different time series. Accordingly, the magnitude can differ and one may prefer to first calculate a key figure on test set level and then aggregate it. However, due to the large values of hyperparameter, it would have to be the more robust median instead of the mean. Then, one calculates the median of median hyperparameters values. This is less intuitive to interpret and, moreover, the results are fairly similar nonetheless.

the unity constraint can not be fulfilled, see again Equation (4.23). For comparison, we consider the largest weight for every observation in the test set over all time series of the original, unconstrained forecast combination method, i.e., OW. The mean and median weights of the largest OW weights are 16 and 118, respectively. The distribution of the largest OW weights for each test set is right-skewed (15.5 skewness) and for one observation, the largest weight was about 23279. For 95% of the observations in all test sets the maximum weight is greater two. Accordingly, $\gamma_0 = 1.1$ constraints weights substantially. This holds similar for $L_1(1/N)$ and $L_1(\dot{\omega})$. Additional evidence that the respective shrinkage parameters actually constrains the weights is the superior forecast accuracy of the considered L_1 methods in comparison to OW.

For the sake of completeness we also include the coefficient of variation to give an indication of the underlying distribution. With respect to the coefficient of variation, the shrinkage parameter of $L_1(0)$ has the least variation, i.e., is most consistent.

Bounded Weights methods Table 7.5 shows the median lower and upper bound of the Bounded Weights methods. Note that for BW($\cdot$) and LB the lower bound has negative values and, as a result, the coefficient of variation is also negative. Nevertheless, it is the absolute deviation from zero that is of interest. Recall, that for BW($\dot{\boldsymbol{\omega}}$) we defined a lower bound deviation from the prior weights. It is a non-negative number that is subtracted from prior weights.

	BW($\cdot$)		BW($\boldsymbol{\omega}^{E2}$)		BW($\boldsymbol{\omega}^{E3}$)		BW($\boldsymbol{\omega}^{IL}$)		LB
	$\underline{\omega}$	$\overline{\omega}$	$\underline{\omega}$	$\overline{\omega}$	$\underline{\omega}$	$\overline{\omega}$	$\underline{\omega}$	$\overline{\omega}$	$\underline{\omega}$
Median	0.00	0.40	0.10	0.10	0.10	0.10	0.10	0.10	0.01
CoefVar	-2.69	1.82	2.58	2.52	2.58	2.53	2.57	2.47	-3.16

Table 7.5. Bounded Weights methods: Median and coefficient of variation (CoefVar) of the lower and upper bounds $\underline{\omega}$ and $\overline{\omega}$ for the considered M4 time series.

For the methods that shrink towards prior weights, BW($\dot{\boldsymbol{\omega}}$), both the median lower bound and upper bound deviation is 0.1. Accordingly, for half of the test observations across all time series, a small deviation from prior weights leads to a better forecast accuracy. For about 40% of test set observations for BW($\dot{\boldsymbol{\omega}}$) a deviation of zero is estimated to be the best hyperparameter value, i.e., the prior weights where used to combine forecasts.

For LB the median lower bound is positive. This is evidence that the extension of feasible values of the lower bound we proposed is substantial to improve forecast accuracy, at least in the validation sets that were used for hyperparameter estimation. For BW($\cdot$) the lower bound is zero or greater for half of the test set observations, i.e., only non-negative weights are considered. The median upper bound is roughly three to four times the equal weights value ($1/9$), i.e., the solution most likely contains one or

two sets of identically weighted forecasts and individual weights, recall Section 5.1.2. With a lower bound of zero, one of the identically weighted sets can have zero weights, which means that the methods performs a variable selection.

With respect to the coefficient of variation, there is more variation for the estimated values of the lower than the upper bound. The largest variation is given for LB. The BW($\dot{\boldsymbol{\omega}}$) methods have a similar a variation of the estimated hyperparameters.

Individual Feature Bounds methods The Individual Feature Bounds methods have many parameters that determine the individual feature lower and upper bounds. Recall, that because we are also using prior weights, we determine not the bound themselves but again the deviation from prior weights. Those deviations are within the individual deviation (IFD) vectors for the lower and upper bounds (IFB). The lower and upper IFD vectors are $\underline{\boldsymbol{\aleph}}$ and $\overline{\boldsymbol{\aleph}}$, respectively.

As discussed in Section 6.2, we limited the number of candidate values of the hyperparameters as follows:

- smallest deviation: $\psi_{min} \in \{0, 0.1, 0.2, 0.3\}$, see Section 6.1.1.
- largest deviation:$\psi_{max} \in \{\psi_{min} + 0.1, \psi_{min} + 0.2, \psi_{min} + 0.3\}$, see Section 6.1.1.
- transformation function: $\Psi \in \{step, GRelU\}$, see Sections 6.1.2.2 and 6.1.2.3.
- feature: $\xi() \in \{MSE, AvgMSEC, AccDiv\}$, see Equations (2.2), (6.46) and (6.47).
- feature values: $\boldsymbol{\nu} \in \{ranked, not\ ranked\}$, see Section 6.1.1.

For the lower and upper IFDs separately, we allow for different smallest and largest deviation (ψ_{min}, ψ_{max}). For *GRelU* we use a threshold $\check{\nu} = 0.5$ and for *step* we use $\Upsilon = 2$ steps, see again Section 6.1.4.

For each time series we compute the median value of ψ_{min}, ψ_{max} for both $\underline{\boldsymbol{\aleph}}$ and $\overline{\boldsymbol{\aleph}}$ for all test set observations of the 989 time series. The result are in Table 7.6.

		IFB(·)	IFB($\dot{\omega}$)		
			$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
$\underline{\boldsymbol{\aleph}}$	ψ_{min}	0.00	0.00	0.00	0.00
	ψ_{max}	0.10	0.20	0.20	0.20
$\overline{\boldsymbol{\aleph}}$	ψ_{min}	0.00	0.00	0.00	0.00
	ψ_{max}	0.30	0.20	0.20	0.20

Table 7.6. Individual Feature Bounds methods: Median value for the smallest and largest deviation for both $\underline{\boldsymbol{\aleph}}$ and $\overline{\boldsymbol{\aleph}}$ (IFD) for the considered M4 time series.

For all methods and both $\underline{\boldsymbol{\aleph}}$ and $\overline{\boldsymbol{\aleph}}$ the smallest deviation ψ_{min} has a median value of zero, i.e., it is zero for half of all observations in the test sets. In that case the *step*

and *GRelU* select at least one forecast to have zero weight, if the feature values are not ranked. Otherwise, about half of the forecasts have zero weights given the parameter values we chose for the transformation functions ($\check{\nu} = 0.5$ and $\Upsilon = 2$). Accordingly, the variable selection property of the transformation function were beneficial in the validation sets and, thus, were chosen for forecasting.

ψ_{max} has the same median value for both the lower and upper deviation for all IFB($\dot{\omega}$). For IFB($\cdot$) negative weights are overall more constrained than positive weights as ψ_{max} is smaller for $\underline{\aleph}$ than for $\overline{\aleph}$.

With respect to the transformation functions, for about 60% of observations in the test sets the *step* function is used for both IFB($\cdot$) and IFB($\dot{\omega}$). Forecasts with a feature value smaller or equal to 0.5 where constraint to have zero weights. This corresponds to half of the forecasts, if ranked features values are used. Otherwise, the proportion can vary, because the scaled feature values are not distributed equally in the feature vector, recall Section 6.1.1. Correspondingly, for roughly 40% the *GRelU* provided the best results in the validation set. Recall, that *GRelU* also constrains weights to zero ,as described above for *step*, for feature values smaller or equal to $\check{\nu} = 0.5$. However, for those 40% of time series and observations in the test sets it was beneficial to allow for a linear increase in the IFD for forecast with a feature value greater 0.5, instead of a step change as for *step*.

Table 7.7 shows the percentages of observations across all test sets for which a particular feature is used by the Individual Feature Bounds methods. The Individual Feature

	IFB($\cdot$)	IFB($\dot{\omega}$)		
		$\boldsymbol{\omega}^{E2}$	$\boldsymbol{\omega}^{E3}$	$\boldsymbol{\omega}^{IL}$
MSE	0.61	0.51	0.51	0.50
AvgMSEC	0.14	0.20	0.20	0.21
AccDiv	0.25	0.29	0.29	0.30

Table 7.7. Individual Feature Bounds methods: Percentage of how often each method uses the features MSE, AvgMSEC and AccDiv for each time series and the corresponding test set observations.

Bounds methods that impose bounds around prior weights (IFB($\dot{\omega}$)) use the MSE as a feature for about half of the observations in the test sets. The amount of (average) diversity between forecasts is in fact useful to determine bounds for about a fifth of observations. Consequently, in 30% of the cases the average of the scaled MSE and AvgMSEC is the most useful feature. In comparison, for IFB($\cdot$) the MSE is used more often for about ten percentages points. Both AvgMSEC and AccDiv are used less for about five percentage points of time series and observations.

Overall, this provides strong evidence that the diversity of forecast is an important and powerful feature for forecast combination with Individual Feature Bounds. Either by itself (AvgMSEC) or together with the forecast accuracy (AccDiv).

With respect to the feature values for all IFB methods roughly 60% of observations in the test sets used ranked feature values while, correspondingly, 40% used the calculated, scaled feature values. It shows that both of these approaches add value and are preferable for certain situations or for specific time series.

In summary, there is no one true hyperparameter setup that works for every time series. The complexity of time series and forecasts leads to structures that require different constraints to combine weights such that the best forecast accuracy can be achieved. One should always evaluate various setups of hyperparameters to find the method and hyperparameter pair that is best for a given situation.

7.2.3 Forecast Accuracy with Respect to Forecast Features and Characteristics

Table 7.8 shows the median relMSE of each method for a set of time series. We segmented the time series into four groups with respect to the mean error correlation of input forecasts. To this end, we used the $25, 50$ and 75%-quantiles to sort time series into the four groups. Accordingly, a time series where the mean error correlation of forecasts is between the 25% and 50% quantile is part of the second group, i.e., column two. The mean error correlation of each group is depicted in the first row. We use the mean error correlation as an indicator of the overall magnitude of the error correlation matrix, because, in contrast to the simulation study, real world error correlation matrices are very different to one another and follow a less organized structure.

Overall, one can see that the median relMSE increases for all methods the larger the mean error correlation is, at least in comparison to the EW forecast. It has to be noted that, because we use the relMSE, changes between groups is an interaction of changes in EW and the corresponding method. Because the main goal of this chapter was to evaluate the forecast combination methods overall, we used the easily interpretable relMSE instead of something like the MASE, where forecasts are scaled relative to the in-sample naive forecast, see again Section 6.1.3.1. Accordingly, if the relMSE increase, it basically means that the difference between the forecast capability of EW and the corresponding forecast combination method decreases. With respect to this, considering the overall increase in relMSE for larger mean error correlation is curious. Overall, we saw in the simulation studies and know theoretically, that the EW is best for smaller or no error correlation. The given results, however, lead to the conclusion that the forecast combination methods potentially have a worse forecast accuracy for larger error correlations. We observed this already for the MSE values in the simulation study, see for example CM1 to CM3 in Table 4.1. The higher the error correlation, the larger the MSE values are. Although, if we calculate the relMSE for the corresponding scenarios

Mean Error Cor		0.44	0.56	0.63	0.69
EW		1.00	1.00	1.00	1.00
PW		0.65	0.77	0.84	0.97
OW		2.49	3.35	12.58	28.61
IL		0.64	0.72	0.83	0.95
LHS	LS	2.49	3.35	12.58	28.61
	LHS	0.63	**0.68**	**0.79**	**0.93**
$L_1(\kappa)$	0	0.70	0.79	0.87	1.00
	$1/N$	0.71	0.81	0.91	1.00
$L_1(\dot{\boldsymbol{\omega}})$	$\boldsymbol{\omega}^{E2}$	0.63	0.71	0.81	0.96
	$\boldsymbol{\omega}^{E3}$	0.64	0.77	0.89	0.99
	$\boldsymbol{\omega}^{IL}$	0.64	0.74	0.84	0.96
BW($\cdot$)		0.72	0.80	0.92	0.99
BW($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	0.62	0.72	0.82	0.94
	$\boldsymbol{\omega}^{E3}$	0.64	0.72	0.84	0.96
	$\boldsymbol{\omega}^{IL}$	0.62	0.75	0.85	0.95
LB		0.73	0.82	0.92	1.00
IFB($\cdot$)		0.65	0.73	0.84	0.97
IFB($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	0.62	0.69	0.82	**0.93**
	$\boldsymbol{\omega}^{E3}$	**0.61**	0.71	0.82	0.94
	$\boldsymbol{\omega}^{IL}$	0.62	0.71	0.83	0.95

Table 7.8. relMSE with respect to the mean error correlation segmented in four groups.

MSE CoefVar		0.45	0.95	1.52	2.23
EW		1.00	1.00	1.00	1.00
PW		1.00	0.79	0.77	0.67
OW		15.53	2.77	6.69	5.09
IL		**0.98**	0.79	0.68	**0.55**
LHS	LS	15.53	2.77	6.69	5.09
	LHS	0.99	**0.73**	0.68	**0.55**
$L_1(\kappa)$	0	1.03	0.81	0.79	0.72
	$1/N$	1.01	0.85	0.81	0.70
$L_1(\dot{\boldsymbol{\omega}})$	$\boldsymbol{\omega}^{E2}$	1.01	0.76	0.69	0.57
	$\boldsymbol{\omega}^{E3}$	1.00	0.78	0.75	0.73
	$\boldsymbol{\omega}^{IL}$	1.01	0.77	0.71	0.58
BW($\cdot$)		1.01	0.84	0.79	0.65
BW($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	1.01	0.76	0.70	0.56
	$\boldsymbol{\omega}^{E3}$	0.99	0.77	0.69	0.56
	$\boldsymbol{\omega}^{IL}$	1.00	0.79	0.71	0.57
LB		1.01	0.83	0.82	0.71
IFB($\cdot$)		1.03	0.77	0.72	0.57
IFB($\dot{\omega}$)	$\boldsymbol{\omega}^{E2}$	1.00	0.78	**0.66**	0.56
	$\boldsymbol{\omega}^{E3}$	1.00	0.77	0.68	**0.55**
	$\boldsymbol{\omega}^{IL}$	1.00	0.79	0.69	0.56

Table 7.9. relMSE with respect to the MSE coefficient of variation segmented in four groups.

over CM1 to CM3, it gets smaller the higher the correlation is. This showcases how complex and different real-world time series are as there are many factors that influence the forecast accuracy.

Let us now consider the relMSE values of specific forecast combination methods (Table 7.8). For the benchmark methods only PW and IL have competitive median relMSE values. EW and OW as well as LS have noticeably larger median relMSE values than all other methods for all four segments. The IFB methods with prior weights have the smallest median relMSE (0.61) for the group with the smallest mean error correlation. With respect to the last group, IFB($\dot{\omega}$) and LHS both have the smallest values (0.93). For the groups in-between, LHS is better (0.68 and 0.79). Oftentimes, however, the distance between IFB($\dot{\omega}$) and LHS is small (except for group three), e.g., group two LHS has 0.68 and IFB($\dot{\omega}$) have 0.69 to 0.71. This holds similarly for the other methods, if they shrink towards prior weights, i.e., $L_1(\dot{\boldsymbol{\omega}})$ and BW($\dot{\omega}$).

Table 7.9 presents the median relMSE grouped by the coefficient of variation of the MSE. This refers to the error variance similarity that we used for the simulation study, see Chapter 3. The larger the coefficient of variation, the large the standard deviation relative to the mean, i.e., the more diverse the MSE values of input forecasts are.

For more similar MSE values of forecast, the equal weights is the though benchmark that it has been throughout the years. Most of the forecast combination methods have

Number of Obs.		34	61	77	125
EW		1.00	1.00	1.00	1.00
PW		0.97	0.79	0.79	0.82
OW		23.62	6.69	2.53	6.30
IL		**0.84**	0.73	0.74	0.83
LHS	LS	23.62	6.69	2.53	6.30
	LHS	**0.84**	0.70	**0.73**	0.79
$L_1(\kappa)$	0	1.02	0.79	0.81	0.82
	$1/N$	1.00	0.82	0.80	0.84
$L_1(\dot{\boldsymbol{\omega}})$	$\boldsymbol{\omega}^{E2}$	0.87	0.71	0.76	**0.78**
	$\boldsymbol{\omega}^{E3}$	0.91	0.77	0.77	0.87
	$\boldsymbol{\omega}^{IL}$	0.89	0.74	0.76	0.81
BW$(\cdot)$		1.00	0.82	0.79	0.83
BW$(\dot{\omega})$	$\boldsymbol{\omega}^{E2}$	0.90	0.71	0.75	0.80
	$\boldsymbol{\omega}^{E3}$	0.91	0.70	0.76	0.79
	$\boldsymbol{\omega}^{IL}$	0.91	0.72	0.75	0.81
LB		1.00	0.84	0.81	0.84
IFB$(\cdot)$		0.89	0.76	0.76	0.81
IFB$(\dot{\omega})$	$\boldsymbol{\omega}^{E2}$	0.87	0.70	0.77	0.80
	$\boldsymbol{\omega}^{E3}$	0.86	**0.69**	0.76	0.79
	$\boldsymbol{\omega}^{IL}$	0.89	0.72	0.76	0.81

Table 7.10. relMSE with respect to the number of training observations segmented in four groups.

relMSE values greater one, i.e., the MSE of EW is smaller. The only exceptions are IL (0.98), LHS and BW($\boldsymbol{\omega}^{E2}$) (0.99). This result is to an extent similar to what one expects theoretically and what we observed in the simulation study. EW is theoretically optimal if all forecasts have the same error variance and no error correlation. In the simulation study we also saw that, EW was better the more similar the forecast error variance were, see for example Table 4.11 and Figure 4.16.

The more diverse the forecast are in terms of their forecast accuracy the smaller the median relMSE of all methods gets, as we would expect. For example, IFB($\dot{\omega}$) have a relMSE of about one for the smallest coefficient of variation, and then it decreases down to 0.55 and 0.56 for the group with the most dissimilar forecasts in terms of their accuracy.

For groups two, three and four, both the forecast combination methods that shrink towards prior weights and LHS have competitive forecasts, i.e., they have similarly small relMSE values. The only exception is $L_1(\boldsymbol{\omega}^{E3})$ with a relMSE of 0.73 for group four while all other methods that shrink towards prior weights have values between 0.55 and 0.58. With respect to the methods that do not shrink towards prior weights IFB$(\cdot)$ has smaller, more competitive relMSE values than BW$(\cdot)$ and $L_1(\kappa)$ for all groups except the first one with the smallest coefficient of variation.

Table 7.10 shows the median relMSE of the forecast combination methods segmented by the number of observations. The average number of observations for each segment is depicted in the first row.

As we expected, the relMSE of all forecast combination methods first decreases from group one to group two (median number of observation 34 and 61) as more observations are available both for weight and hyperparameter estimation. Thereby, the estimation uncertainty decreases. However, the relMSE does not further decrease for group three (77) and even increases again for group four (125). In particular the increase from groups three to four may be evidence that an increasing number of training observation is not always the best choice. At its core, this highlights the difference in the ideas of an expanding and a rolling window forecast procedure. The expanding window reduces estimation uncertainty as the estimation is based on more observations. In contrast, the rolling window can react faster and more effectively to shot-term structures in the data. With respect to Table 7.10, this benefit of a rolling window may be mitigated by a too large window, i.e., to many observations are considered in the training. The increase in median relMSE from group three to four is smallest for $L_1(0)$.

Table 7.10 shows that for more observations, the shrinkage methods, in particular the methods we proposed BW($\dot{\omega}$) and IFB($\dot{\omega}$) have similar median relMSE to the best method. See for example, IFB($\dot{\omega}$) for group two and four or BW($\dot{\omega}$) for group two and three. These findings hold similar for LHS. If the number of observations is limited, $L_1(\dot{\omega})$ and IFB($\dot{\omega}$) are the most promising shrinkage methods, however, LHS has the smallest median relMSE.

7.3 Summary

The overall result of the 989 monthly M4 time series highlight the value of the new forecast combination methods we proposed. They are a valuable complement to existing approaches.

Shrinkage methods in general and the methods we proposed in particular (BW and IFB) strictly improve out-of-sample forecast accuracy compared to the four benchmarks methods for the majority of time series (74% and 60% respectively). This holds particularly for our proposed new field of forecast combination where we constrain weights individually based on feature values (IFB). Compared to all other methods, it is the single best method most often (22.24%). Nonetheless, BW also is superior for about 103 (10.41%) time series.

With respect to the hyperparameters we observed that overall the shrinkage parameters and bounds are more tight in comparison to how large the unconstrained weights of the OW solution are at times. However, there is variation between observations or time series. This is particularly well demonstrated when considering the Individual

Feature Bound methods as they have a lot of hyperparameters which all provided value for specific observations or time series.

With respect to forecast features or characteristics we can conclude that real-world data can be much more complex than simulation studies. Although, we would expect that the forecast combination methods are better in comparison to EW the higher the mean error correlation is, our results suggest the opposite. However, because we have real-world data many aspects can influence the forecast accuracy of a method simultaneously. In contrast, for the monthly M4 time series we observed that the more dissimilar the nine forecasts are in terms of their MSE, the better sophisticated forecast combination methods are compared to EW. This is in line with our observations within the simulation study. With respect to the number of time series, the results indicate that at first an increase in the number of observation in the training set has a positive effect on the forecast accuracy. However, if there are too many observations in the training set, the relMSE can again increase. This can be due to the fact that with such a large training set, the forecast combination methods can not appropriately capture short-term structures in the data.

Overall, the methods we proposed, in particular forecast combination with individual feature bounds around prior weights, are a valuable addition to existing approaches. They should be taken into consideration for forecast combination problems. Bounded Weights and Individual Feature Bounds have the single best forecast accuracy for 39% of the time series compared to all other methods, i.e., Benchmarks, LHS, and L_1.

For real-world applications, it is strongly advisable to utilize all methods, benchmarks and more sophisticated methods, and evaluate their performance in a validation set to identify the superior method for a given dataset. This can lead to a far superior forecasting performance compared to choosing one single method or hyperparameter set up for different or even the same time series at various times.

8 Conclusion

The original forecast combination problem by J. M. Bates and Granger (1969) estimates weights by minimizing the forecast error variance of the combined forecast subject to the unity constraint. Forecast combination allows us to combine forecasts that are based on different information and models or experts, thereby creating a more accurate and robust forecast through diversification. Over the years many forecast combination methods have been developed. However, for over half a century, simple combination approaches have frequently outperformed more sophisticated or theoretically optimal methods — a phenomenon commonly referred to as the forecast combination puzzle. Evidence suggests that this is due to estimation error and the sensitivity of weights. We introduced and discussed all the above, i.e., forecasting and forecast combination, the forecast combination puzzle in Chapter 2 and presented a brief literature overview.

The overarching objective of this thesis is summarized in our first research question: *how to further improve the forecast accuracy of a combined forecast using constrained weights?* Accordingly, in this thesis we consider the forecast combination problem with constrained weights. The constraint we impose shrinks weights towards a shrinkage direction or prior weights. We can improve the forecast accuracy by allowing for a small bias which leads to a reduction in the error variance of the combined forecast.

In order to analyze forecast combination methods within this thesis, in Chapter 3 we first analyzed forecasts from the Europeans Centrals Banks Survey of Professional Forecasters to gain insights into the structure of real-world forecast combination data. Based on these insights and the simulation study of Roccazzella et al. (2022), we developed an extended framework for simulation studies. This framework is intended to be used for analyzing forecast combination methods. With this simulation study framework we can analyze the methods with respect to the error correlation structure of forecasts, the similarity of the forecast accuracy or error variance respectively, and groups of forecast with noticeably smaller or larger error variances. Furthermore, we designed the simulation study such that it can easily be extended, e.g., allowing for the incorporation of different distributions of error variances between forecasts or varying error correlations within groups of forecasts (Contribution I). In this thesis we analyzed all considered forecast combination methods using this simulation study.

In Chapter 4 we showed how shrinkage can improve forecast accuracy. Allowing for a small bias can lead to a reduction in variance. We also introduced the concept of using an L_1 constraint for shrinkage in forecast combination. Based on this, we devel-

oped a unified framework that is based on the original forecast combination problem by J. M. Bates and Granger (1969), namely a quadratic optimization problem that minimizes the error variance subject to linear constraints including the unity constraint. The unified framework allows for a comprehensive comparison of all considered L_1 methods on the same basis (Contribution II). This includes shrinkage towards fixed values such as zero or the equal weights forecast as well as shrinkage towards prior weights. Inspired by other forecast combination methods like the egalitarian, partially egalitarian Lasso, and the Linear Hybrid Shrinkage, we defined *Conditional Group Equal Weights* as a shrinkage direction for forecast combination with constrained weights (Contribution III).

As a result of the L_1 constraint, some forecasts are omitted from the combination. Moreover, the L_1 constraints restricts the weight vector as a whole and, allowing a single forecast to contribute significantly to the combined forecast in comparison to other forecasts. Although, both a selection of forecasts and the large contribution of a single forecast leads to an improved forecast accuracy, to an extent it is contrary direction to an idea of forecast combination: diversification. Forecast combination is intended to create robust combined forecasts by incorporating multiple forecasts. In Chapter 5 we introduced a method that leads to a more diversified forecast while also shrinking weights: *Forecast Combination with Bounded Weights* (Contribution IV). Instead of applying a single constraint for the whole weight vector, we propose using both common lower and upper bounds or constraints for each weight individually. With the addition of the upper bounds, we create a method that nests competitive benchmark methods (Positive Weights, Equal Weights) and the original forecast combination problem (Optimal Weights). Furthermore, we extend our approach to utilize bounds around prior weights, which function as a shrinkage direction upon we improve.

Based on the Bounded Weights method, we propose a new forecast combination method and thereby research direction: *Forecast Combination with Individual Feature Bounds* in Chapter 6 (Contribution V). This method follows the concept of the original forecast combination problem described by J. M. Bates and Granger (1969). It defines individual lower and upper bounds that are based on feature values of the forecasts. The incorporation of feature values allows for the utilization of additional external information, which can be specifically chosen for the requisite forecast application. Additionally, the use of different transformation functions enables us to further influence the shape of the constraints. We can incorporate various properties for the individual feature bounds, such as a variable selection based on the feature values if many forecasts are given. Accordingly, we can design our method in many ways, e.g., with and without a variable selection (L_1 versus Bounded Weights) or with a design more similar to common bounds for some forecasts, while others are less constrained due to their favorable feature values. Moreover, the use of Individual Feature Bounds is adaptable

to a multitude of applications and requirements. To illustrate, we can define a feature for underestimation, thereby constraining forecasts that are prone it. Alternatively, if the combined forecast requires greater consistency, that is, only minor changes between time periods, we can define a feature for this and use it to derive Individual Feature Bounds. We can define a feature that measures how fast a forecast adapts to short-term changes such that the combined forecast more quickly captures economic shocks. In retail, we may prefer a forecast that is particularly well-suited to forecast demand throughout promotional periods. Consequently, we can evaluate each input forecast only for those periods and incorporate the results as a feature based on which we derive the individual feature bounds. Both the incorporation of external information tailored towards the application and the properties we can enforce by the choice and design of the transformation function showcase that our method is flexible and holds great potential to improve forecast accuracy across all applications.

We conducted a comprehensive simulation study for all considered forecast combination methods to evaluate their performance. We analyzed how the forecast combination methods perform for scenarios designed with respect to different error correlation matrices, the degree of similarity between the input forecast accuracies and special groups. First and foremost, in each ex post analysis of Chapters 4 to 6 we found that the corresponding forecast combination with constrained weights is in general superior to the benchmark methods, also with respect to groups of scenarios. Another key finding is that using prior weights to shrink towards is a viable approach to combine forecasts for all L_1, Bounded Weights and Individual Feature Bounds. In particular, shrinkage towards prior weights is useful for highly correlated forecast errors and, to a certain extent, if the input forecasts are more dissimilar with respect to their error variance (forecast accuracy, respectively).

A comparison of all forecast combination methods reveals that the Individual Feature Bounds are clearly superior. They have the smallest MSE (among others) for all but two scenarios, i.e., about 97% of scenarios. In comparison, the benchmark methods, Linear Hybrid Shrinkage, L_1 methods and Bounded Weights have the smallest MSE for, respectively, roughly 3%, 3%, 10%, and 10% of scenarios.

However, throughout Chapters 4 to 6 we found that as soon as the hyperparameters of the forecast combination methods have to be estimated a priori (out-of-sample analysis) rather than chosen ex post, the benchmark methods (except the original forecast combination problem: Optimal Weights) are much more competitive in terms of their forecast accuracy. In other words, the estimation of hyperparameters introduces uncertainty, which negates the clear superiority of forecast combination methods with constrained weights. However, in scenarios with highly correlated forecast errors and more dissimilar error variances, forecast combination methods which shrink towards prior weights, in particular Conditional Group Equal Weights, are still competitive.

Nevertheless, in particular Positive Weights (PW) and Inverse-Loss Weights (IL) provide a superior forecast performance more consistently.

The results of the simulation study indicate that the forecast combination methods we considered and proposed in this thesis have great potential to improve the forecast accuracy. However, this potential cannot be fully realized due to the uncertainty associated with the estimation of the hyperparameters. The findings suggest that further analysis of hyperparameter estimation is necessary to fully utilize the potential for improvement in forecast accuracy offered by the forecast combination methods.

In addition to the simulation studies, we conducted a comprehensive empirical analysis using nearly 1000 time series from the M4 competition. In contrast to the simulation study, the considered forecast combination methods (L_1, Linear Hybrid Shrinkage, Bounded Weights and Individual Feature Bounds) provide good result even though hyperparameters were estimated (except for Linear Shrinkage). The L_1 methods and Bounded Weights methods have the strictly best forecast accuracy for about 10% of time series and Linear Hybrid Shrinkage for about 13%. With Individual Feature Bounds we have the strictly best forecast accuracy for 22% of the time series which is more often than the benchmark methods with 20%. If we consider how often each method has the best forecast accuracy among others, both the benchmarks and Individual Feature Bounds are the best method most often for about a third of the time series. For comparison, L_1, Linear Hybrid Shrinkage and Bounded Weights have, among others, the best forecast accuracy for about 23%, 23%, and 25%, respectively.

In general, shrinkage methods and the forecast combination methods with constrained weights we proposed (Bounded Weights and Individual Feature Bounds) strictly improve the forecast accuracy for 74% and 60% of these real-world time series compared to the benchmark methods. If we compare Bounded Weights and Individual Feature Bounds (BW($\cdot$), BW($\dot{\omega}$), IFB($\cdot$), and IFB($\dot{\omega}$)) against all other considered forecast combination methods (benchmarks, Linear Hybrid Shrinkage and L_1), our methods have strictly improved the forecast accuracy for 39% of the real-world time series, i.e., 388 out of 989 time series.

In this thesis, we analyzed the forecast combination methods for both simulated and real-world data. Our analysis offers valuable insights and demonstrates the great potential of forecast combination with constrained weights, in particular Individual Feature Bounds. In particular the empirical analysis allows us to assess the usefulness or value of our newly proposed forecast combination methods for real-word time series. Due to the fact that we use numerous time series, we ensure that our results are meaningful and robust (Contribution VI).

In this thesis and our research, there are limitations or opportunities that need to be considered and that provide the foundation for future research.

First, the results from the simulation study suggest that we need to further investigate the hyperparameter estimation of the forecast combination methods. Although the result for the M4 time series demonstrated that the forecast accuracy was improved relative to the benchmark methods for numerous time series, there is potential for further improvement, which is driven by the hyperparameter estimation. To this end, we can utilize, for instance, the error correlation matrix and error variances of the input forecasts to train a neural network that predicts the hyperparameters.

Second, there are a lot of different forecast combination methods and approaches beyond forecast combination with constrained weights that are important to consider, including double shrinkage via weighted least squares Lasso proposed by Liu et al. (2023). This was, however, beyond the scope of this thesis. A future comprehensive comparison on real-world data could prove invaluable in assessing the utility of both Bounded Weights and Individual Feature Bounds.

Third, our out-of-sample analysis of the M4 time series and the simulation study demonstrated that there is no single forecast combination method that is optimal for all time series or scenarios. Consequently, future research should focus on developing an a priori approach for determining which forecast combination methods to use for a specific time series. To this end, we can use cross-validation, whereby, we choose the forecast combination method with its hyperparameters that has the best forecast accuracy within the validation set. Alternatively, we can develop an algorithm that is based on the insights from our analysis of the forecast accuracies with respect to the error correlation matrix, degree of similarity of forecast accuracy, the number of observations in the training set, and so on.

Fourth, the second research question of this thesis was: *how to incorporate additional, external information in forecast combination with constrained weights?* To this end, we introduced Forecast Combination with Individual Feature Bounds, which achieve this. However, as we discussed in Chapter 6, one can argue that we can use a different objective function to tailor the combined forecast for a particular application. Nevertheless, we do believe that Forecast Combination with Individual Feature Bounds is the best approach. It allows for a more flexible and easily applicable incorporation various external information and by the use of the transformation function, it can enforce properties such as variable selection. However, for future research, we would like to investigate the usefulness of prior weights based on feature values and customized objective to analyze how well additional information is incorporated by those approaches in comparison to Individual Feature Bounds.

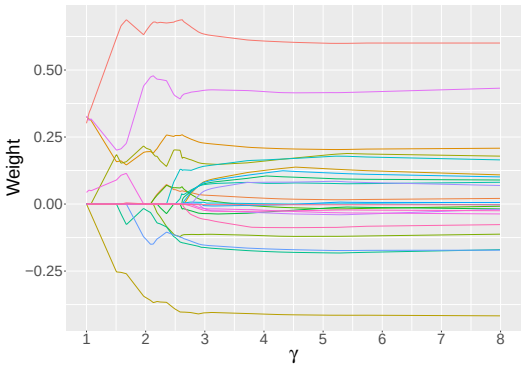
Lastly, a key advantage of Individual Feature Bounds is its capability to incorporate additional, external information thereby enabling to be designed for a specific application. Therefore, for future research we suggest analyzing Individual Feature Bounds across various applications with application-specific features. For instance, in

economics, traffic, energy consumption, energy generation, reverse logistics, interest and exchanges rates, stock returns, electricity prices, climate change, epidemics and pandemics, risk of violence, elections, or sports and many more.

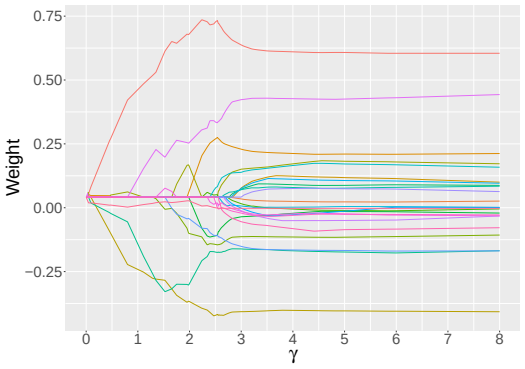
Accurate forecasts are of great importance as they are an indispensable component of today's everyday life for individuals, society, economics, and businesses alike. The overarching research question of this thesis was *how to further improve the forecast accuracy of a combined forecast using constrained weights?* We showed that our proposed methods for combined forecasts using additional constraints within the original forecast combination problem can achieve this objective. While benchmark methods like the equal weights forecast are competitive forecast combination methods not only for an Ox weighting competition in 1906 England but also for more complex real-world time series, we can confidently state that Bounded Weights and Individual Feature Bounds are a valuable addition to the existing forecast combination methods that improve the forecast accuracy of the combined forecast.

A Appendix Chapter 4

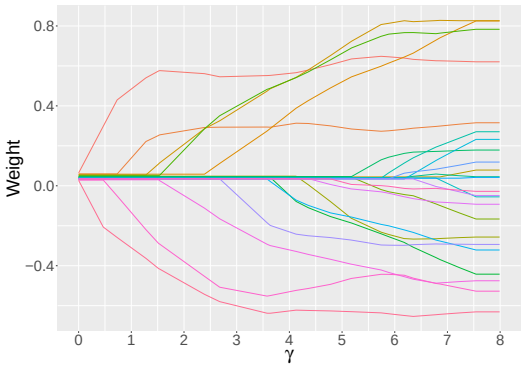
Additional Examples for Weight Paths



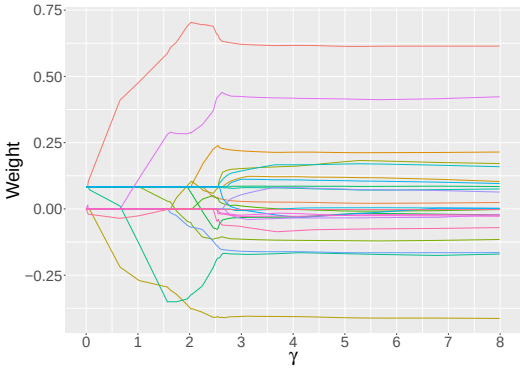
(a) $L_1(0)$



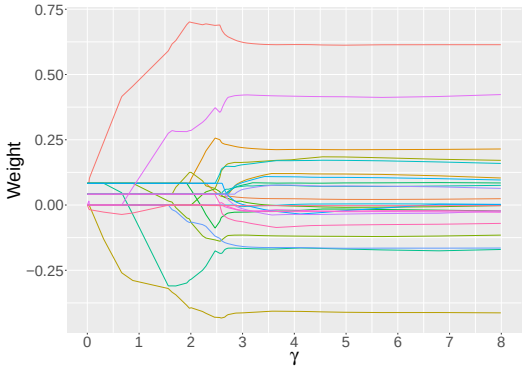
(b) $L_1(1/N)$



(c) $L_1(\boldsymbol{\omega}^{IL})$



(d) $L_1(\boldsymbol{\omega}^{E2})$



(e) $L_1(\boldsymbol{\omega}^{E3})$

Figure A.1. Additional weight paths for the L_1 methods with a different simulated data set with $N = 24$ forecasts for $L_1(0)$,$L_1(1/N)$,$L_1(\boldsymbol{\omega}^{IL})$,$L_1(\boldsymbol{\omega}^{E2})$ and $L_1(\boldsymbol{\omega}^{E3})$. The data was generated using the simulation study presented in Section 3.2 with $CM1$, a relative group distance $z = 0.5$ and no special group.

One-Step Procedure for the Partially Egalitarian Lasso

The partially egalitarian Lasso was discussed in Section 4.1, and it is presented in Equation (4.13) with $q = 1$. The idea of Diebold and Shin (2019) is that some weights are shrunken and selected towards zero, while others are shrunken and selected towards equal weights. The partially egalitarian Lasso with the unity constraint is defined by

$$\begin{aligned} \underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\omega} \\ \text{subject to} \quad & \boldsymbol{\omega}'\mathbf{1} = 1, \\ & \|\boldsymbol{w}\| \leq \gamma_0, \\ & \|\boldsymbol{w} - {}^{1}/_{\|\boldsymbol{\omega}\|_0}\| \leq \gamma_{{}^{1}/_{\|\boldsymbol{\omega}\|_0}} \end{aligned} \tag{A.1}$$

The second constraint restricts the deviation from zero, i.e., it shrinks towards zero. The third constraint restricts the deviation from equal weights of the currently non-zero weights by using the L_0 norm first described in Equation (4.8). In Equation (A.1) we adapted the peLasso for the forecast combination problem with a unity constraint. However, the first constraint can be omitted if preferred.

Diebold and Shin (2019) only presented a two-step procedure due to difficulties to solve Equation (4.13) directly. We can, however, incorporate both constraints simultaneously and present the one-step procedure that was left for future research. To this end, we reformulate the optimization problem from Equation (A.1) similarly we reformulated the optimization problem depicted in Equation (4.28) into the optimization problem shown in Equation (4.32). We introduce two sets of N additional variables u_i and v_i $\forall i = 1 \ldots, N$ for the two L_1 norms. Additionally, for the L_0 norm within the second constraint we introduce N binary variables $b_i \in \{0, 1\}$. We define a new covariance matrix as

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{A.2}$$

where, again, $\mathbf{0}$ is an $N \times N$ matrix that contains only zeros. The weights vector is defined as

$$\tilde{\boldsymbol{\omega}} = (\omega_1, \omega_2, \ldots, \omega_N, u_1, u_2, \ldots, u_N, v_1, v_2, \ldots, v_N, b_1, b_2, \ldots, b_N)'. \tag{A.3}$$

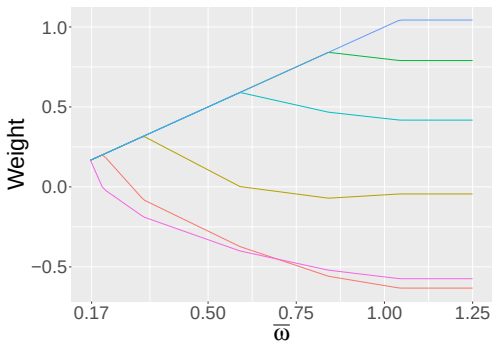
As a result, the corresponding optimization problem is given by

$$
\begin{aligned}
\underset{\boldsymbol{\omega}}{\text{minimize}} \quad & \tilde{\boldsymbol{\omega}}' \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\omega}} \\
\text{subject to} \quad & \sum_{i=1}^{N} \omega_i = 1, \\
& \sum_{i=1}^{N} u_i \leq \gamma_{(1/\sum_{j=1}^{N} b_j)}, \\
& \sum_{i=1}^{N} v_i \leq \gamma_0, \\
\omega_i - 1/\sum_{j=1}^{N} b_j \leq u_i \qquad & \forall\, i = 1, \ldots, N, \\
1/\sum_{j=1}^{N} b_j - \omega_i \leq u_i \qquad & \forall\, i = 1, \ldots, N, \\
\omega_i \leq v_i \qquad & \forall\, i = 1, \ldots, N, \\
-\omega_i \leq v_i \qquad & \forall\, i = 1, \ldots, N, \\
\omega_i \geq \zeta b_i \qquad & \forall\, i = 1, \ldots, N, \\
\omega_i \leq -\zeta b_i \qquad & \forall\, i = 1, \ldots, N, \\
\omega_i, u_i, v_i \in \mathbb{R} \qquad & \forall\, i = 1, \ldots, N, \\
b_i \in \{0, 1\} \qquad & \forall\, i = 1, \ldots, N
\end{aligned}
\tag{A.4}
$$

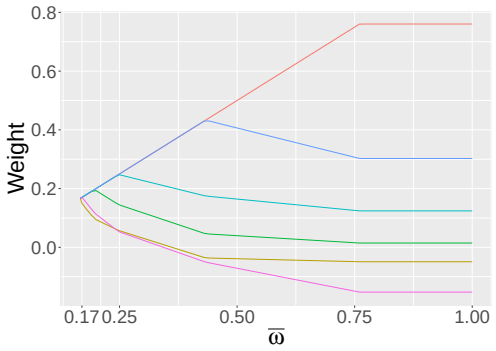
The parameter γ_0 as well as the variables v_i corresponds to the L_1 constraint that shrinks parameters towards zero. The corresponding parameter to the L_1 constraint that restricts the deviations from the equal weights solution is $\gamma_{(1/\sum_{j=1}^{N} b_j)}$ and the variables are u_i. These constraints were introduced within Section 4.2.2.1. In order to model the L_0 constraint linearly, the last two inequalities are used. Given an arbitrary small positive number ζ, the two inequalities force b_i to be one, if the weight is non-zero. To be more precise, if the weights deviation from zero exceeds the threshold ζ (or $-\zeta$). Changing the variables b_i alone has no impact on the objective function. However, if a weight has to have a value smaller (or larger) ζ (or $-\zeta$), b_i has to be set to zero. Thereby, the sum of $b_i \; \forall \; i = 1, \ldots, N$ corresponds the number of non-zero elements. Therefore, using it in the constraints that correspond to $\gamma_{(1/\sum_{j=1}^{N} b_j)}$ and u_i results in a shrinkage towards the equal weights of the currently non-zero elements of $\omega_i \; \forall \; i = 1, \ldots, N$.

The quadratic mixed integer problem of Equation (A.4) introduces a one-step procedure that Diebold and Shin (2019) have left for future research.

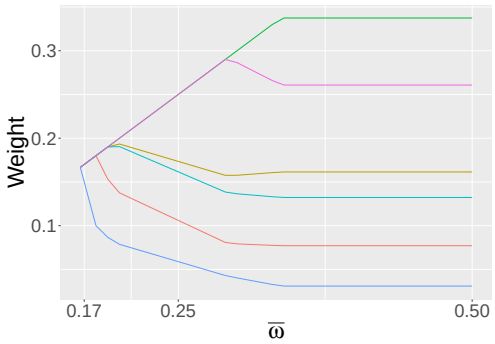
B Appendix Chapter 5



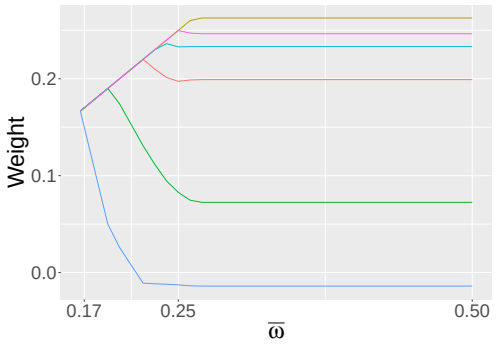
(a) CM1, $z = 0.05$ and SG: none.



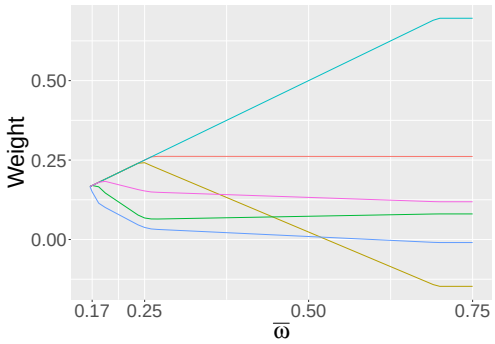
(b) CM2, $z = 0.20$ and SG: first.



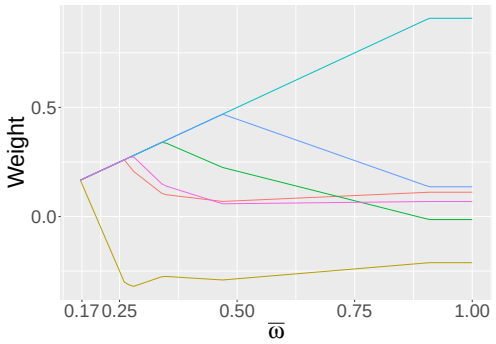
(c) CM3, $z = 0.50$ and SG: last.



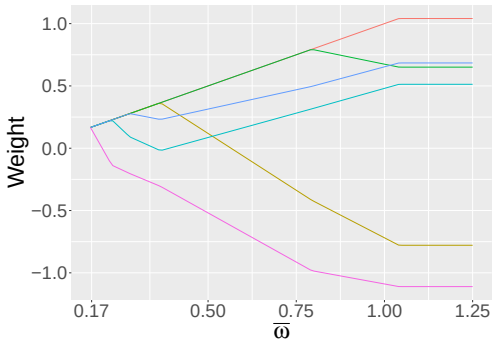
(d) CM4, $z = 0.05$ and SG: first and last.



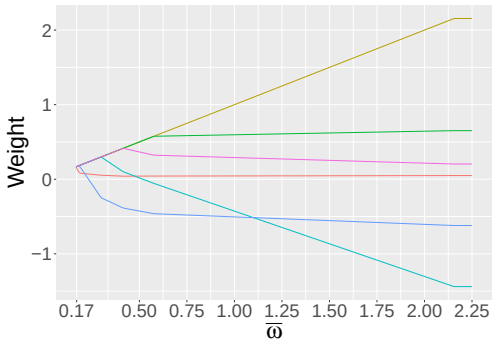
(e) CM5, $z = 0.20$ and SG: none.



(f) CM6, $z = 0.50$ and SG: first.

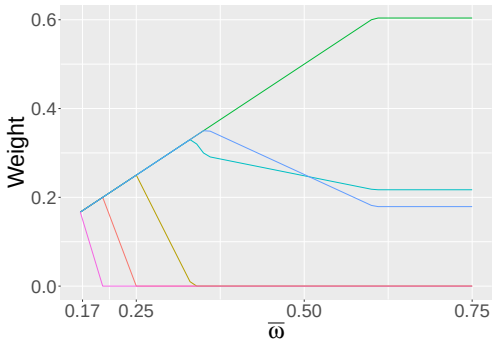


(g) M4 time series one.

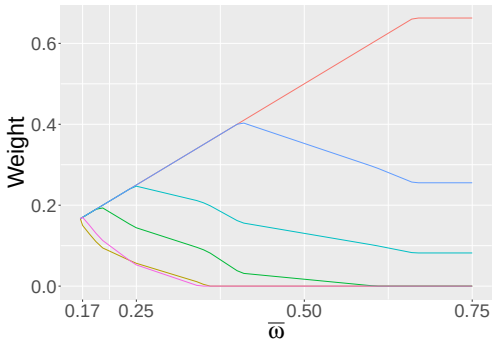


(h) M4 time series two.

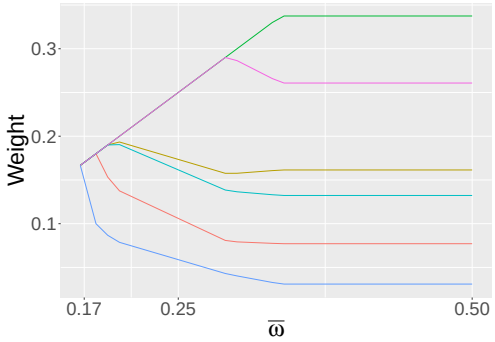
Figure B.1. Examples of forecast weight paths for the transitions $\overleftrightarrow{EO}$. The data was created on the basis of the simulation study of Section 3.2 for the scenario specified in the caption of each sub-figure. Six forecasts where chosen randomly out of 24. The last two sub-figures show two randomly drawn time series from the M4 monthly data set. For the last 30 observations forecasts where computed based on the methods in Chapter 7.



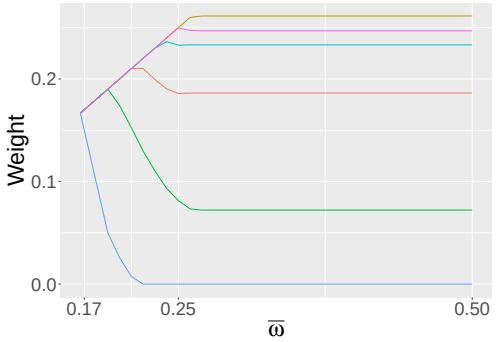
(a) CM1, $z = 0.05$ and SG: none.



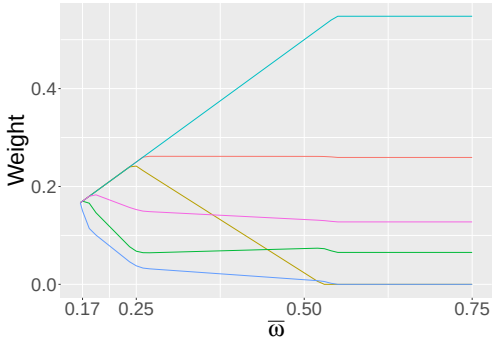
(b) CM2, $z = 0.20$ and SG: first.



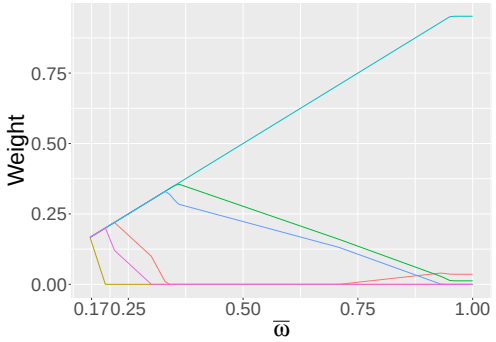
(c) CM3, $z = 0.50$ and SG: last.



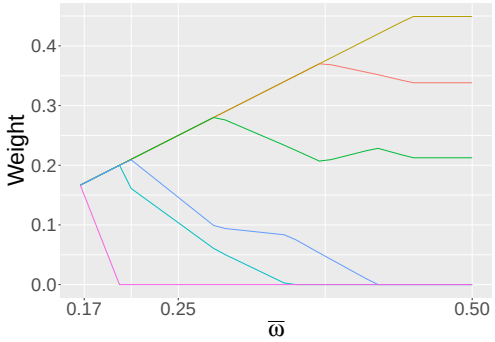
(d) CM4, $z = 0.05$ and SG: first and last.



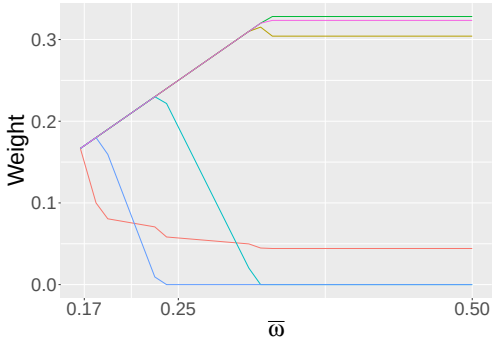
(e) CM5, $z = 0.20$ and SG: none.



(f) CM6, $z = 0.50$ and SG: first.

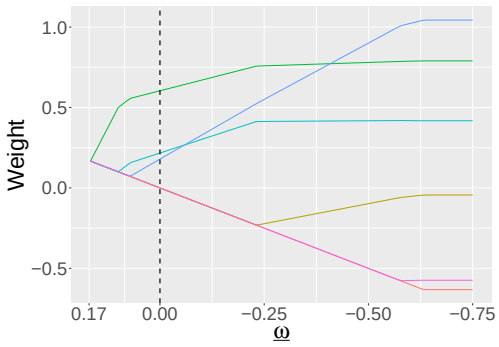


(g) M4 time series one.

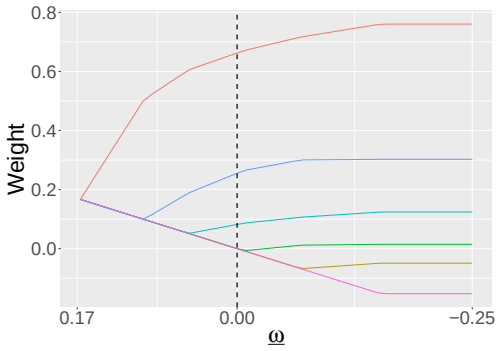


(h) M4 time series two.

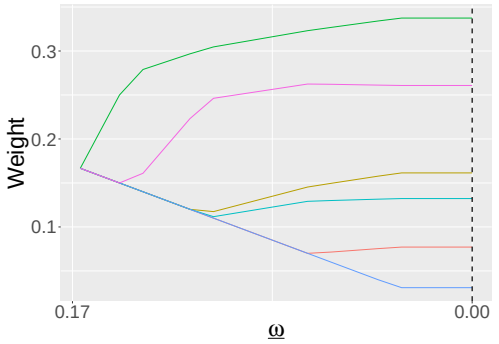
Figure B.2. Examples of forecast weight paths for the transitions $\overleftrightarrow{EP}_{ub}$. The data was created on the basis of the simulation study of Section 3.2 for the scenario specified in the caption of each sub-figure. Six forecasts where chosen randomly out of 24. The last two sub-figures show two randomly drawn time series from the M4 monthly data set. For the last 30 observations forecasts where computed based on the methods in Chapter 7.



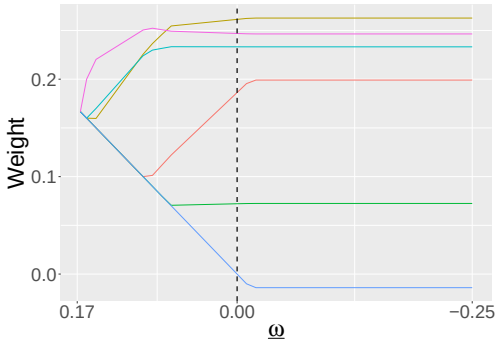
(a) CM1, $z = 0.05$ and SG: none.



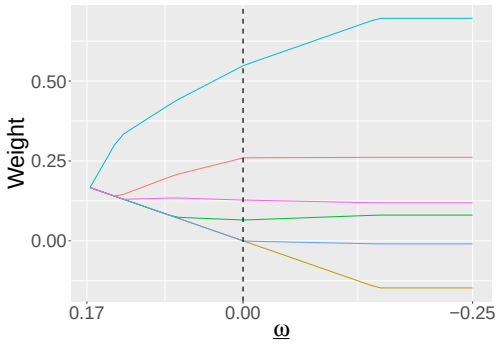
(b) CM2, $z = 0.20$ and SG: first.



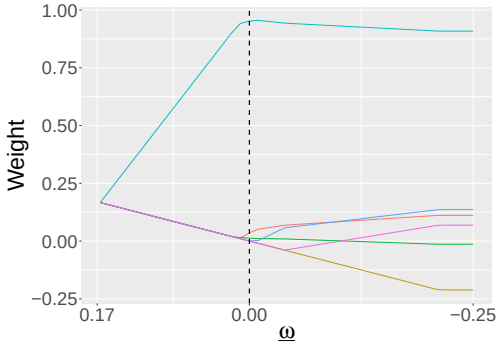
(c) CM3, $z = 0.50$ and SG: last.



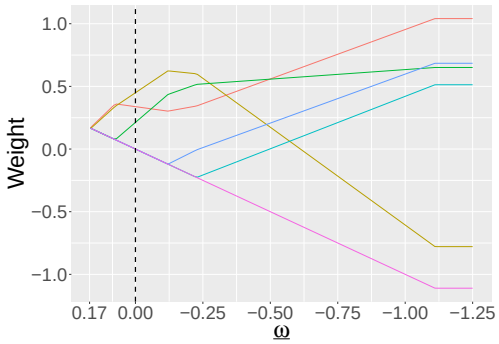
(d) CM4, $z = 0.05$ and SG: first and last.



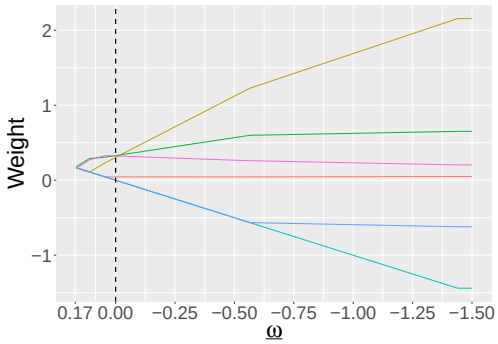
(e) CM5, $z = 0.20$ and SG: none.



(f) CM6, $z = 0.50$ and SG: first.



(g) M4 time series one.



(h) M4 time series two.

Figure B.3. Examples of forecast weight paths for the transitions $\overleftrightarrow{EPO}$. The data was created on the basis of the simulation study of Section 3.2 for the scenario specified in the caption of each sub-figure. Six forecasts where chosen randomly out of 24. The last two sub-figures show two randomly drawn time series from the M4 monthly data set. For the last 30 observations forecasts where computed based on the methods in Chapter 7.

C Appendix Chapter 6

Further Ideas for Unified Accuracy and Diversity Measures

For Individual Feature Bounds we want to use a feature that captures both the forecast accuracy of the candidate forecasts and their diversity. With a similar objective Wang, Kang, Petropoulos, and Li (2022) use a version of the MSE decomposition from Equation (6.42) as a measurement. To this end, they assume equally weighted forecasts. The *"Accuracy-Diversity Trade-off (ADT)"*(Wang, Kang, Petropoulos, & Li, 2022) is given by

$$\begin{aligned} \text{ADT} &= \text{AvgMSE} - \iota \text{AvgMSEC}, && \text{(C.1)} \\ &= \frac{1}{N}\sum_{i=1}^{N} \text{MSE}_i - \iota \frac{1}{N^2}\sum_{i=1}^{N-1}\sum_{j>i}^{N} \text{MSEC}_{i,j}. && \text{(C.2)} \end{aligned}$$

The parameter $\iota \in [0, 1]$ controls to what extent diversity is taken into consideration relative to the accuracy. If $\iota = 0$ the ADT only takes the forecast accuracy into account and, thus, is identical to the average MSE of all forecasts. With a parameter value of $\iota = 1$ the ADT takes diversity into consideration as much possible. Note that for $\iota > 1$ the ADT can become negative. In Wang, Kang, Petropoulos, and Li (2022) the ADT measure is used within their RAD algorithm to evaluate subsets of forecasts. In context of the Individual Feature Bounds we are interested in a similar measure but for a single forecast such that we can use it as a feature. This proves to be more difficult.

Average ADT One potential approach is to artificially assume that the set of forecast only consists of two forecasts. Then we calculate the ADT for each of the $N(N-1)$ pairs of forecasts, i.e.,

$$\text{ADT}_{i,j}(\iota) = \frac{\text{MSE}_i + \text{MSE}_j}{2} - \iota \frac{\text{MSEC}_{i,j}}{4} \qquad \forall\, i, j \in \{1, \ldots, N\} : i \neq j. \qquad \text{(C.3)}$$

After that we determine the average ADT of forecast i by

$$\text{AvgADT}_i(\iota) = \frac{1}{N-1} \sum_{j \in \{1,\ldots,N\}: i \neq j} \text{ADT}_{i,j}(\iota) \qquad \forall\, i = 1, \ldots, N. \qquad \text{(C.4)}$$

Accordingly, we use the average Accuracy-Diversity Trade-off (AvgADT) as our feature values, i.e., $\nu_i = \text{AvgADT}_i$. However, we need to emphasize and address that the feature

values are determined under assumptions and then used in a different environment. First, the $\text{ADT}_{i,j}$ values[58] used in AvgADT_i are calculated within a forecast combination scenario with only two forecast and, second, equal weights are assumed. Thereafter, however, the AvgADT_i is used as a feature to derive constraints for a N forecast scenario. Nevertheless, we argue that the AvgADT_i can still be a beneficial feature to use. It gives an indication of how a certain forecast performs in combination with other forecasts and how different forecasts are based on the choice of ι. Moreover, the equal weights approach has more than once proven itself and contradicted all expectations, see again Section 2.3.

Average Weighted ADT We can build another measurement that considers both accuracy and diversity that does not require the assumption of equal weights. To this end, we simply use the combined MSE decomposition from Equation (6.42) but include the parameter ι. Similar to ADT we create the *Weighted Accuracy-Diversity Trade-Off (WADT)* and define it as

$$\text{WADT}_{i,j}(\iota) = \sum_{i=1}^{N} \omega_i \text{MSE}_i - \iota \sum_{i=1}^{N-1} \sum_{j>i}^{N} \omega_i \omega_j \text{MSEC}_{i,j}. \tag{C.5}$$

Based on this, we can follow the same procedure, assume that $N = 2$ and calculate the WADT for all $N(N-1)$ pairs of forecasts, i.e.,

$$\text{WADT}_{i,j}(\iota) = \omega_i \text{MSE}_i + \omega_j \text{MSE}_j - \iota \omega_i \omega_j \text{MSEC}_{i,j}. \tag{C.6}$$

Again, in the next step we determine the *average WADT (AvgWADT)* of forecast i as

$$\text{AvgWADT}_i(\iota) = \frac{1}{N-1} \sum_{j \in \{1,\dots,N\}: i \neq j} \text{WADT}_{i,j}(\iota) \qquad \forall\, i = 1, \dots, N. \tag{C.7}$$

Of course, for the AvgWADT_i we need to determine weights, see Equations (C.5) and (C.6) respectively. If we use equal weights it simplifies to AvgADT_i. However, we have a similar problem as before. We use a measurement for the accuracy and diversity that is derived within a two forecast scenario. Under the same conditions the weights are determined. Then we use this measurement as features to derive individual feature bounds for a different scenario ($N > 2$). Of course, within this framework it is not possible to use the same weights that we derive to impose the constraints and, thus, we need to choose the weights for the AvgWADT.

It is not sensible to use the optimal weights from the original N forecasts scenario problem instead. In that case the weights of two forecasts i and j in Equations (C.5) and (C.6) respectively do not sum to one. As a result, the first part of Equation (C.5)

[58] Note that we omit ι when referring to $\text{ADT}_{i,j}(\iota)$ or $\text{AvgADT}_i(\iota)$ solely for a better readability.

is no longer the average accuracy of the considered forecasts. Instead, the pairwise optimal weights from Equation (C.6), i.e., the optimal weights calculated for each pair of forecasts $(i, j) \forall\ i, j \in \{1, \ldots, N\} : i \neq j$ can be used.

List of Symbols

$\mathbf{0}$	$N \times 1$ vector of zeros
$\mathbf{1}$	$N \times 1$ vector of ones
$\boldsymbol{\alpha}$	$S \times 1$ vector of relative difference between the median error variance of any group
$\boldsymbol{\aleph}$	individual feature deviation vector - allowed deviation from prior weights that define the individual lower and upper bound.
$\underline{\boldsymbol{\aleph}}, \overline{\boldsymbol{\aleph}}$	lower, upper individual feature deviation vector - allowed deviation from prior weights that define the individual lower and upper bound.
AvgMSEC_i	average MSEC between forecast i and any other forecast
$\underline{\aleph}, \overline{\aleph}$	lower, upper individual feature deviation - allowed deviation from prior weights that define the individual lower and upper bound.
$\underline{\aleph}^*, \overline{\aleph}^*$	adjusted lower, upper individual feature deviation to ensure the feasibility of the optimization problem.
$\overline{\beth}^*(\underline{\beth})$	maximum upper bound deviation - the solution will not change for any $\overline{\beth} > \overline{\beth}^*(\underline{\beth})$
$\text{BW}(\cdot)$	bounded prior weights without prior weights (forecast combination method)
$\text{BW}(\dot{\omega})$	bounded around prior weights weights / bounded prior weights (forecast combination method)
$\underline{\beth}, \overline{\beth}$	lower, upper bound deviation from prior weights
$\underline{\beth}^*(\overline{\beth})$	minimum lower bound deviation - the solution will not change for any $\underline{\beth} < \underline{\beth}^*(\overline{\beth})$
BW	bounded weights (forecast combination method)
c	index for the combined forecast
CGEW	conditional group equal weights
CM	correlation matrix
$d(\boldsymbol{\omega}, \dot{\boldsymbol{\omega}})$	measure of diversion between the weight vector $\boldsymbol{\omega}$ and a reference prior weight vector $\dot{\boldsymbol{\omega}}$
$\boldsymbol{\eta}$	$S \times 1$ dimensional vectors or median error variances.

$\boldsymbol{\varepsilon}$	$N \times 1$ vector of forecast errors
ϵ	regression error term
η_s	median error variances
$\overleftrightarrow{EPO}$	path or transition between EW, PW and OW
$\overleftrightarrow{EP}_{lb}$	path or transition between EW and PW by varying the lower bound
$\overleftrightarrow{EP}_{ub}$	path or transition between EW and PW by varying the upper bound
ε	forecast error
ECB	european central bank
eLasso	egalitarian lest absolute shrinkage and selection operator
EW	equal weights (forecast combination method)
FFORMA	feature-based forecast model averaging
γ	shrinkage parameter or intensity
γ^*	value of shrinkage parameter or intensity at which it does not constrain or effect the solution
$\gamma_\kappa, \gamma_{\dot{\boldsymbol{\omega}}}$	shrinkage parameter or intensity towards the shrinkage direction κ or $\dot{\boldsymbol{\omega}}$ respectively
$\mathbb{G}_i$	group i of forecasts
G	number of groups for the CGEW
g_j	budget of group j of the solution for the smallest feasible γ.
GDP	gross domestic product
GRelU	generalized rectified linear unit
h	forecast horizon
HICP	harmonized Index of Consumer Prices inflation rate
IFB($\cdot$)	individual feature bounds without prior weights (forecast combination method)
IFB($\dot{\omega}$)	individual feature bounds around prior weights (forecast combination method)
IFB	individual feature bounds (forecast combination method)
IL	inverse-loss weighted average (forecast combination method)
IQR	interquartile range
κ	shrinkage direction
λ	shrinkage parameter or intensity

λ^*	value of shrinkage parameter or intensity at which it does not constrain or effect the solution
$L_1(\kappa)$	shrinkage towards a fixed value κ via L_1 constraints (forecast combination method)
IL	inverse-loss weighted average (forecast combination method)
Lasso	lest absolute shrinkage and selection operator
LB	lower bound (forecast combination method)
LHS	linear hybrid shrinkage
LS	linear shrinkage
MAE	mean absolute error
MAPE	mean absolute percentage error
MASE	mean absolute scaled error
MSE	mean squared error
MSEC	mean squared error for coherence
relMSE	relative mean squared error - MSE of a method divided by a reference MSE (EW) to scale it
$\boldsymbol{\nu}$	feature vector of forecasts
$\check{\boldsymbol{\nu}}$	another scaled feature value for the generalized logistic function to ensure the desired smallest and largest deviations from prior weights
$\check{\nu}$	threshold for the feature value ν within the GRelU
$\| \ \|_q$	L_q-norm
ν_i	feature value of forecasts i
$\tilde{\boldsymbol{\nu}}$	scaled feature vector of forecasts
$\tilde{\nu}_i$	scaled feature value of forecasts i
N	number of forecasts
$\boldsymbol{\omega}^{LHS}$	forecast combination weight of lhs
$\boldsymbol{\omega}$	$N \times 1$ vector of forecast combination weights
$\boldsymbol{\omega}^{E2}$	prior weights vector and shrinkage direction with two groups of conditional equal weight
$\boldsymbol{\omega}^{E3}$	prior weights vector and shrinkage direction with two groups of conditional equal weight
$\boldsymbol{\omega}^{IL}$	prior weights vector and shrinkage direction inverse-loss weighted average

$\dot{\boldsymbol{\omega}}$	prior weight vector
ω	forecast combination weight
ω^{OW}	forecast combination weight for optimal weights
$\overline{\omega}^*(\underline{\omega})$	maximum upper bound - the solution will not change for any $\overline{\omega} > \overline{\omega}^*(\underline{\omega})$
$\underline{\omega}, \overline{\omega}$	lower, upper bound for forecast combination weights
$\underline{\omega}^*(\overline{\omega})$	minimum lower bound - the solution will not change for any $\underline{\omega} < \underline{\omega}^*(\overline{\omega})$
$\widehat{\boldsymbol{\omega}}^{LHS}(\lambda)$	estimated forecast combination weight of lhs given λ
OLS	ordinary least squares estimator
OW	optimal weights (forecast combination method)
$\check{\psi}_{min}, \check{\psi}_{max}$	adjusted smallest, largest individual feature deviation from prior weights for the generalized logistic function to ensure the desired smallest and largest deviations from prior weights
$\overleftrightarrow{PO}$	path or transition between PW and OW
ϕ_1, ϕ_2, ϕ_3	parameters that influence the generalized logistic function
Ψ	transformation function that maps an input to the individual feature bound
ψ_{min}, ψ_{max}	smallest, largest individual feature deviation from prior weights
peLasso	partially egalitarian lest absolute shrinkage and selection operator
PW	positive weights (forecast combination method)
$\bar{\rho}^{r,s}$	common error correlation between any two forecasts $i, j = 1, \ldots, N$ of any two groups $s, r = 1, \ldots, S$
$\rho_{i,j}^{s,r}$	error correlation between any two forecasts $i, j = 1, \ldots, N$ of any two groups $s, r = 1, \ldots, S$
$\rho_{i,j}$	error correlation of forecasts i and j
$\boldsymbol{\Sigma}$	$N \times N$ error variance covariance matrix
$\boldsymbol{\sigma}_s^2$	$N/S \times 1$ dimensional vectors or error variances of group s.
$\hat{\sigma}_c^2(\boldsymbol{\omega})$	empirical error variance
$\sigma_c^2(\boldsymbol{\omega})$	actual error variance
$\sigma_{i,j}^2$	error covariance of forecasts i and j
σ_i, σ_i^2	error standard deviation, error variance of forecast i
$\widehat{\boldsymbol{\Sigma}}$	$N \times N$ estimated error variance covariance matrix

s, S	index and number of forecast groups within the simulation study
SG	special groups
SPF	survey of Professional Forecasters
τ	Total number of observations in a training set of a time series
T	Total number of observations of a time series
t	time index
$\check{\Upsilon}$	step size of the step transformation function given Υ
Υ	number of steps in the step transformation function
u_i, v_i	auxiliary variables
$\boldsymbol{v}$	$N \times 1$ vector to select forecasts to shrink towards zero or EW (LHS method)
$\xi(\ )$	function that calculates the feature values
$\hat{\boldsymbol{y}}$	$N \times 1$ vector of forecasts
$\hat{y}_{c,t}$	combined forecast for time t
$\hat{y}_{i,t}$	forecast i for time t
y_t	observation of a time series at time t
ζ	arbitrary small positive number
z_s	error variance similarity - relative distance of the median error variance of group s compared to group one

References

Aggarwal, C. C. (2023). *Neural networks and deep learning: A textbook* (2nd ed.). Cham: Springer International Publishing. doi: 10.1007/978-3-031-29642-0

Aksu, C., & Gunter, S. I. (1992). An empirical analysis of the accuracy of sa, ols, erls and nrls combination forecasts. *International Journal of Forecasting*, *8*(1), 27–43. doi: 10.1016/0169-2070(92)90005-T

Arratia, A. (2014). *Computational finance: An introductory course with r* (Vol. 1). Paris: Atlantis Press. doi: 10.2991/978-94-6239-070-6

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530. doi: 10.1016/s0169-2070(00)00066-2

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197–200. doi: 10.1016/j.ijforecast.2019.03.010

Babikir, A., & Mwambi, H. (2016). Evaluating the combined forecasts of the dynamic factor model and the artificial neural network model using linear and nonlinear combining methods. *Empirical Economics*, *51*(4), 1541–1556. doi: 10.1007/s00181-015-1049-1

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*(1), 68–75. doi: 10.1287/mnsc.41.1.68

Bates, D., Maechler, M., & Jagan, M. (2022). *Matrix: Sparse and dense matrix classes and methods: R package version 1.5-1.* Retrieved from `https://CRAN.R-project.org/package=Matrix`

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. doi: 10.1057/jors.1969.103

Blanc, S. M., & Setzer, T. (2020). Bias–variance trade-off and shrinkage of weights in forecast combination. *Management Science*, *66*(12), 5720–5737. doi: 10.1287/mnsc.2019.3476

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), 587–603. doi: 10.1016/j.ijforecast.2020.07.007

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., & Rautanen, T. (2007). The ecb survey of professional forecasters (spf) - a review after eight years' experience. *ECB occasional paper*(59).

Box, G. E. P., Luceño, A., & Del Paniagua-Quinones, M. C. (2009). *Statistical control*

by monitoring and adjustment (2nd ed., Vol. v.700). Hoboken, NJ: Wiley.

Capistrán, C., & Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, *27*(4), 428–440. doi: 10.1198/jbes.2009.07211

Causton, D. R. (1969). A computer program for fitting the richards function. *Biometrics*, *25*(2), 401–409. doi: 10.2307/2528797

Chan, F., & Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, *34*(1), 64–74. doi: 10.1016/j.ijforecast.2017.08.005

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In B. Krishnapuram (Ed.), *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY: ACM. doi: 10.1145/2939672.2939785

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762. doi: 10.1016/j.ijforecast.2015.12.005

Clemen, R. T. (1986). Linear constraints and the efficiency of combined forecasts. *Journal of Forecasting*, *5*(1), 31–38. doi: 10.1002/for.3980050104

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. doi: 10.1016/0169-2070(89)90012-5

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, *6*(1), 3–73.

Conflitti, C., de Mol, C., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, *31*(4), 1096–1103. doi: 10.1016/j.ijforecast.2015.03.009

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130–143. doi: 10.1287/deca.2015.0315

de Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513–1527. doi: 10.1198/jasa.2011.tm09771

Dickinson, J. P. (1975). Some comments on the combination of forecasts. *Journal of the Operational Research Society*, *26*(1), 205–210. doi: 10.1057/jors.1975.43

Diebold, F. X., & Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, *6*(4), 503–508. doi: 10.1016/0169-2070(90)90028-A

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal*

of Forecasting, *35*(4), 1679–1691. doi: 10.1016/j.ijforecast.2018.09.006

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*(426), 463–475. doi: 10.1080/01621459.1994.10476768

Elliott, G., & Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, *122*(1), 47–79. doi: 10.1016/j.jeconom.2003.10.019

Elliott, G., & Timmermann, A. (2016). *Economic forecasting*. Princeton and Oxford: Princeton University Press.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der weg zur datenanalyse* (8th ed.). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fan, J., Zhang, J., & Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, *107*(498), 592–606. doi: 10.1080/01621459.2012.682825

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

Garcia, J. A. (2003). An introduction to the ecb's survey of professional forecasters. *ECB occasional paper*(8).

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. doi: 10.1016/j.ijforecast.2012.06.004

Gentle, J. E. (2017). *Matrix algebra: Theory, computations and applications in statistics* (2nd ed.). Cham: Springer. doi: 10.1007/978-3-319-64867-5

Goldfarb, D., & Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical analysis* (pp. 226–239). Springer Berlin Heidelberg. doi: 10.1007/bfb0092976

Goldfarb, D., & Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, *27*(1), 1–33. doi: 10.1007/BF02591962

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43–54. doi: 10.1016/j.ijforecast.2013.02.005

Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*(2), 197–204. doi: 10.1002/for.3980030207

Gunter, S. I. (1992). Nonnegativity restricted least squares combinations. *International Journal of Forecasting*, *8*(1), 45–59. doi: 10.1016/0169-2070(92)90006-U

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*(1), 1–13. doi: 10.1016/j.ijforecast.2006.08.001

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Higham, N. J. (2002). Computing the nearest correlation matrix–a problem from

finance. *IMA Journal of Numerical Analysis*, *22*(3), 329–343. doi: 10.1093/imanum/22.3.329

Hurlbert, G. H. (2010). *Linear optimization: The simplex workbook.* New York and Heidelberg: Springer. doi: 10.1007/978-0-387-79148-7

Hyndman, R. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7–14. doi: 10.1016/j.ijforecast.2019.03.015

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). Melbourne, Australia: Otexts.com/ffp3. Accessed on 16.12.2023.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models: R package version 8.21.* Retrieved from `https://pkg.robjhyndman.com/forecast/`

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, *27*(3), 1–22. doi: 10.18637/jss.v027.i03

Hyndman, R., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. doi: 10.1016/j.ijforecast.2006.03.001

Inoue, A., Jin, L., & Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, *196*(1), 55–67. doi: 10.1016/j.jeconom.2016.03.006

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. E. (2023). *An introduction to statistical learning: With applications in python.* Cham: Springer International Publishing. doi: 10.1007/978-3-031-38747-0

James, W., & Stein, C. (1992). Estimation with quadratic loss. In S. Kotz (Ed.), *Breakthroughs in statistics* (pp. 443–460). New York, NY: Springer New York. doi: 10.1007/978-1-4612-0919-5_30

Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, *207*. doi: 10.1016/j.eswa.2022.117921

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*(1), 163–169. doi: 10.1016/j.ijforecast.2007.06.001

Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2022). Forecast with forecasts: Diversity matters. *European Journal of Operational Research*, *301*(1), 180–190. doi: 10.1016/j.ejor.2021.10.024

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. doi: 10.1016/j.ijforecast.2015.12.003

Kolassa, S. (2011). Combining exponential smoothing forecasts using akaike weights.

International Journal of Forecasting, *27*(2), 238–251. doi: 10.1016/j.ijforecast.2010.04.006

Krasnopolsky, V. M., & Lin, Y. (2012). A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental us. *Advances in Meteorology*, *2012*, 1–11. doi: 10.1155/2012/649450

Krogh, A., & Vedelsby, J. (1994). Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*(7).

Lanza, A., Morigi, S., Selesnick, I. W., & Sgallari, F. (2023). Convex non-convex variational models. In K. Chen, C.-B. Schönlieb, X.-C. Tai, & L. Younes (Eds.), *Handbook of mathematical models and algorithms in computer vision and imaging*. Cham, Switzerland: Springer International Publishing.

Ledoit, O., & Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, *30*(12), 4349–4388. doi: 10.1093/rfs/hhx052

Li, L., Kang, Y., & Li, F. (2023). Bayesian forecast combination using time-varying features. *International Journal of Forecasting*, *39*(3), 1287–1302. doi: 10.1016/j.ijforecast.2022.06.002

Li, L., Kang, Y., Petropoulos, F., & Li, F. (2023). Feature-based intermittent demand forecast combinations: accuracy and inventory implications. *International Journal of Production Research*, *61*(22), 7557–7572. doi: 10.1080/00207543.2022.2153941

Lichtendahl, K. C., & Winkler, R. L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting*, *36*(1), 142–149. doi: 10.1016/j.ijforecast.2019.03.027

Liu, L., Hao, X., & Wang, Y. (2023). Solving the forecast combination puzzle using double shrinkages*. *Oxford Bulletin of Economics and Statistics*. doi: 10.1111/obes.12590

Luenberger, D. G., & Ye, Y. (2016). *Linear and nonlinear programming* (Fourth edition ed., Vol. volume 228). Cham: Springer. doi: 10.1007/978-3-319-18842-3

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., ... Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153. doi: 10.1002/for.3980010202

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The m2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–22. doi: 10.1016/0169-2070(93)90044-n

Makridakis, S., & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. doi: 10.1016/s0169-2070(00)00057-1

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808. doi: 10.1016/j.ijforecast.2018.06.001

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. doi: 10.1016/j.ijforecast.2019.04.014

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*(4), 1325–1336. doi: 10.1016/j.ijforecast.2021.07.007

Makridakis, S., Spiliotis, E., Hollyman, R., Petropoulos, F., Swanson, N., & Gaba, A. (2023). The m6 forecasting competition: Bridging the gap between forecasting and investment decisions. *arXiv preprint arXiv:2310.13357*, *2023*.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. doi: 10.1037/a0036677

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77. doi: 10.2307/2975974

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., & Ramírez-Hassan, A. (2022). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*, *38*(1), 384–406. doi: 10.1016/j.ijforecast.2021.05.008

Matsypura, D., Thompson, R., & Vasnev, A. L. (2018). Optimal selection of expert forecasts with integer programming. *Omega*, *78*, 165–175. doi: 10.1016/j.omega.2017.06.010

Merkle, E. C., Saw, G., & Davis-Stober, C. (2020). Beating the average forecast: Regularization based on forecaster attributes. *Journal of Mathematical Psychology*, *98*, 102419. doi: 10.1016/j.jmp.2020.102419

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). Fforma: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. doi: 10.1016/j.ijforecast.2019.02.011

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, *137*(2), 131–146. doi: 10.2307/2344546

Newbold, P., & Harvey, D. I. (2008). Forecast combination and encompassing. In W. C. Cockerham (Ed.), *The blackwell companion to medical sociology* (pp. 268–283). Hoboken: Wiley. doi: 10.1002/9780470996430.ch12

Nowotarski, J., Raviv, E., Trück, S., & Weron, R. (2014). An empirical comparison of alternative schemes for combining electricity spot price forecasts. *Energy Economics*, *46*, 395–412. doi: 10.1016/j.eneco.2014.07.014

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben

Taieb, S., ... Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.

Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting*, *36*(1), 110–115. doi: 10.1016/j.ijforecast .2019.01.006

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `https://www.R-project.org/`

Radchenko, P., Vasnev, A. L., & Wang, W. (2023). Too similar to combine? on negative weights in forecast combination. *International Journal of Forecasting*, *39*(1), 18–38. doi: 10.1016/j.ijforecast.2021.08.002

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, *10*(2), 290–301.

Roccazzella, F., Gambetti, P., & Vrins, F. (2022). Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, *38*(1), 97–116. doi: 10.1016/j.ijforecast.2021.04.002

Schmidt, M. (2005). Least squares optimization with l1-norm regularization. *CS542B Project Report*(504), 195–221.

Schulz, F., Setzer, T., & Balla, N. (2022). Linear hybrid shrinkage of weights for forecast selection and combination. *HICSS 2022 : Hawaii International Conference on System Sciences*.

Shaub, D. (2020). Fast and accurate yearly time series forecasting with forecast combinations. *International Journal of Forecasting*, *36*(1), 116–120. doi: 10.1016/j.ijforecast.2019.03.032

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355. doi: 10.1111/ j.1468-0084.2008.00541.x

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the third berkeley symposium on mathematical statistics and probability* (pp. 197–206). Berkeley, CA: University of California Press. doi: 10.1525/9780520313880-018

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. doi: 10.1002/ for.928

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. doi: 10.1016/s0169 -2070(00)00065-0

Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, *35*(2), 474–484. doi: 10.1016/j.ijforecast.2018.10.006

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tsukuma, H., & Kubokawa, T. (2020). *Shrinkage estimation for mean and covariance matrices*. Singapore: Springer Nature Singapore and Imprint Springer. doi: 10 .1007/978-981-15-1596-5

Turlach, B. A., Weingessel, A., & Moler, C. (2019). *quadprog: Functions to solve quadratic programming problems: R package version 1.5-8.* Retrieved from `https://CRAN.R-project.org/package=quadprog`

Tyson, N. d., Strauss, M. A., & Gott, J. R. (2016). *Welcome to the universe: An astrophysical tour*. Princeton: Princeton University Press.

Wallis, K. F. (2014). Revisiting francis galton's forecasting competition. *Statistical Science*, *29*(3). doi: 10.1214/14-sts468

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518–1547. doi: 10.1016/j.ijforecast.2022.11.005

Wang, X., Kang, Y., & Li, F. (2022). Another look at forecast trimming for combinations: robustness, accuracy and diversity. *arXiv:2208.00139*.

Wang, X., Kang, Y., Petropoulos, F., & Li, F. (2022). The uncertainty estimation of feature-based forecast combinations. *Journal of the Operational Research Society*, *73*(5), 979–993. doi: 10.1080/01605682.2021.1880297

Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operations Research*, *40*(3), 609–614. doi: 10.1287/opre.40.3.609