Bias–Variance Aware Integration of Judgmental Forecasts and Statistical Models

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

(Dr.-Ing.)

von der Fakultät für
Wirtschaftswissenschaften
am Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Inform.-Wirt. Sebastian M. Blanc

Tag der mündlichen Prüfung:	19.09.2016
Referent:	Prof. Dr. Thomas Setzer
Korreferent:	Prof. Dr. Gerhard Satzger

Karlsruhe, 2016

Acknowledgements

I would like to express my special gratitude to my advisor Prof. Dr. Thomas Setzer for his great support and trust. He enabled me to pursue this research and motivated me to aim at a deeper understanding of the challenges. His enthusiasm has been an inspiration for my research as well as for my personal development.

I would also like to acknowledge Prof. Dr. Gerhard Satzger, Prof. Dr. Christof Weinhardt, and Prof. Dr. Oliver Grothe, who served on my examination committee and who provided valuable comments on this thesis.

My sincere thanks go to my colleagues in the research group Corporate Services and Systems at the Institute of Information Systems and Marketing and at the FZI Forschungszentrum Informatik. The coffee breaks, discussions, and all the other welcomed distractions greatly contributed to my motivation for writing this thesis.

This work was accomplished in collaboration with the Bayer AG. I thank Bayer –especially Alexander Burck, Kati Schnürer, and the team of the corporate financial controlling– for the financial support and the fruitful cooperation as well as for providing their profound domain knowledge, expertise, and feedback on this research.

Lastly, but most importantly, I thank Yasmin and my family. Yasmin continuously encouraged and supported me with loving care and patience. My family, especially my parents Peter and Karin, were always there for me and believed in me. Without them, this thesis would never have been written.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AP	Agricultural Products
APE	Absolute Percentage Error
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
CoV	Coefficient of Variation
DIV	Diverse
DJ	Dow Jones
DTES	Damped Trend Exponential Smoothing
HP	Health and Pharmaceuticals
IM	Industrial Materials
LAD	Least Absolute Deviation
MA	Moving Average
MSE	Mean Squared Error
OLS	Ordinary Least Squares
OW	Optimal Weights
RMSE	Root Mean Squared Error
RQ	Research Question
RSS	Residual Sum of Squares
SA	Simple Average
sMAPE	Symmetric Mean Absolute Percentage Error
STL	Seasonal Decomposition of Time Series by Loess
WLS	Weighted Least Squares

Part I

Foundations

Chapter 1

Introduction and Motivation

DECISION processes in corporations today often rely on quantitative forecasts, which allow decision makers to anticipate future developments and to take appropriate actions. In corporate planning, forecasts are an important basis for the generation and assessment of alternative courses of action (Hogarth and Makridakis, 1981). In practice, a variety of external and internal forecasts are used in different corporate decisions processes. Important external forecasts are for instance macroeconomic forecasts that are a basis for an assessment of the status of the economy. Sales forecasts are an example of internal forecasts, which are used for business functions such as marketing, production planning, or procurement. Efficiency and efficacy of these business functions strongly depend on the accuracy of the forecasts; inaccurate forecasts can lead to suboptimal decisions and undesirable outcomes.

As corporations are part of a complex and dynamic environment, judgment has proven to be essential for the forecasting of various time series (Lawrence et al., 2006), especially when contextual information plays a major role (Edmundson et al., 1988). Research suggests that even simple eyeballing often leads to credible forecasts (Lawrence et al., 1985). It is consequently not surprising that judgment is widespread in forecasting. Webby and O'Connor (1996) for instance found that judgment is involved in 40-50 % of time series forecasting tasks. More recent studies confirm that corporations still rely on qualitative, judgment-based methods despite the availability of quantitative methods (Klassen and Flores, 2001; Sanders and Manrodt, 2003; McCarthy et al., 2006).

Research from psychology as well as empirical studies of judgmental forecasts show that cognitive biases and heuristics influence forecasts, which are consequently likely to be biased (Hogarth and Makridakis, 1981; Lawrence et al., 2006). As biases in forecasts can decrease accuracy and negatively influence corporate functions (Leitner and Leopold-Wildburger, 2011), approaches aiming at reducing biases and improving forecast accuracy have been developed.

1.1 Motivation

In contrast to human judgment, statistical methods, which in general detect and extrapolate systematic patterns from past data, can be considered objective. For instance, as an alternative to judgmental forecasts, statistical model-based time series forecasting models can be used if a sufficient history of past observations is available. The judgmental and model-based approaches are very different in terms of strengths and weaknesses (Webby and O'Connor, 1996; Makridakis, 1988; Sanders and Manrodt, 2003). While statistical methods can objectively identify patterns in past data, human experts tend to falsely identify signals and non-existing patterns in noisy time series. However, human experts can incorporate qualitative and contextual information that cannot be integrated into statistical forecasting models in a straightforward way. The two approaches are consequently largely complementary in terms of strengths and weaknesses.

One approach aiming at using the strengths of both approaches is a linear combination of forecasts, where the forecasts are combined in a weighted average. Forecast combination is known to increase forecast accuracy (Clemen, 1989) as errors of individual forecasts are compensated, which results in a reduced error variance. The reduction is largest if forecasts are sufficiently diverse regarding error patterns. Judgmental and model-based forecasts are likely to be very different, which makes this combination particularly promising.

Forecast correction, as another approach, assumes that cognitive biases and heuristics as well as their influence on judgmental forecasts are largely stable over time. Statistical methods can then be used to detect systematic linear biases in past judgmental forecasts, which can then be removed from future forecasts.

Both integration mechanisms –forecast correction and combination– require learning a model, which in turns involves estimating parameters that not only fit past forecasts well but also perform well on future forecasts. In forecast combination, the parameters are the weights of the forecasts whereas parameters reflect the biases in judgmental forecasts in forecast correction. In order to find parameters that are well suited for future unknown data, two sources of uncertainty must be considered.

First, the true relationships and thus the optimal parameters of the model are unknown in practice and have to be estimated. The estimation uses an available sample of past observations, which consists of realizations of a random variable from a statistical point of view, and which can by pure coincidence systematically differ from future observations. This is especially the case if small samples are used for the estimation.

The aspect of the estimation uncertainty is reflected by the so-called bias–

variance trade-off from statistical learning theory (Hastie et al., 2009), which is based upon a decomposition of the error when applying a model to unknown data. The error is decomposed into a random error component and two components related to the fit to the data (bias component) and the oversensitivity to the data (variance component). The bias component reflects errors resulting from the estimated parameters, in expectation, differing from the optimal parameters. In contrast, the variance component covers errors resulting from parameter estimates differing between samples from the same population.

Imagine, for instance a very simple model without any parameters is chosen. Certainly, applying the model results in a systematic error as the available training data is not used (strong bias component), but the model does not differ between different training samples from the same population (no variance component). If, in contrast, a complex model is estimated that strongly uses the available training data, the estimated parameters in expectation match the optimal values rather well (low bias component). However, errors are likely to result from differences between training data and future observations (high variance component). This example also illustrates that a trade-off between the bias and the variance component of the error exists. As an optimization of one component increases the other component, a minimization of the overall error requires considering both components.

Second, the relationships and parameters that are estimated might change as a result of shifts or structural changes in the relationship. If, for instance, the expert producing a judgmental forecast changes because of staff turnover, the biases in the judgmental forecast are also likely to change. However, forecast correction or combination models learned from past data are still calibrated to the old error patterns. As a result, a corrected or combined forecast can perform worse than the original judgmental forecast as new errors are introduced.

Overall, while forecast correction and combination are in general promising approaches, the uncertainty resulting in the variance component has to be considered for a successful application in practice. However, established approaches only address the bias component of the error and fit the parameters very closely to available training data. As a consequence, little guidance exists on how to correct or combine judgmental forecasts considering the different sources of uncertainty. Analyzing and understanding forecast correction and combination in terms of the bias–variance trade-off and regarding structural changes is an essential basis for new models with reliable performance on unknown data.

1.2 Research Questions

As uncertainty resulting from parameter estimation and from potential structural breaks influence forecast correction and combination methods, this thesis focuses on developing insights regarding the strength of these influences and on providing guidance on how to correct and combine forecasts. For this purpose, the research questions derived and defined in this section are addressed. Initially, the research questions regarding forecast correction are introduced, followed by those addressing forecast combination.

As the parameters of established linear forecast correction models are estimated from past observations, the size of the available training sample influences the uncertainty of the parameter estimates. Small training samples result in very uncertain estimates and thus a high variance component of the error of the corrected forecast. Depending on the extent of the reduction of the bias component of the error, the strong variance component can outweigh the reduction and result in an overall increase of the error in comparison to the original forecast. As the variance component of the error decreases with increasing training sample size, a minimal training sample size in many cases exists for which the increased variance component resulting from using a forecast correction model is lower than the reduction of the bias component. This minimal training sample size indicates how many observations are required to make it reasonable to apply a forecast correction model. This aspect is addressed in the first research question, RQ 1.

RQ 1 **Forecast Correction – Training Sample Size**
What is the training sample size required so that the variance component of the error of a linearly corrected judgmental forecast is smaller than the reduction of the bias component?

Besides estimation uncertainty, structural breaks can increase the errors when applying forecast correction to unknown data. Structural breaks between past and future observations result in a systematic error, i.e., bias component, that is higher than expected. However, the errors resulting from the estimation uncertainty (i.e. the variance component) remain unchanged. As the bias component is not as strongly reduced as expected, the reduction can in some cases be too low to balance the variance component of the error resulting from applying the correction model. While this effect is likely to be small for weak structural changes, the strength of the effect increases with the strength of structural changes. Thus, structural changes, which are at least as strong as a certain level, result in the corrected forecast having higher overall error than the original forecast. Applying a forecast correction model calibrated using outdated identified linear biases can for instance result in reinforcing the new biases after a structural change.

As a consequence, a relevant basis for the assessment of the robustness of forecast correction methods is to quantify how large structural changes can be for the corrected forecast to still outperform the original one. If structural changes have substantial influence, considering potential breaks in the parameter estimation procedure may be of importance for successful application of forecast correction methods. Different approaches can in principle be used for this purpose that might however increase estimation uncertainty resulting in increased errors. On the one hand, structural breaks can, with considerable uncertainty, be detected in past data. If a structural break is identified, it can be considered in the estimation of the parameters of forecast correction models. However, because of the uncertainty in the detection of structural breaks, some structural changes are likely to be diagnosed falsely. On the other hand, a weighting of past data can be introduced, which reduces the influence of old observations with potentially outdated biases. While the influence of the old biases fades out if structural changes exist, the uncertainty of the estimates is increased because of the stronger fit to few recent observations. It is however unknown whether including structural changes is beneficial and, if so, which approach is more beneficial. These aspects are formulated in RQ 2.

RQ 2 Forecast Correction – Structural Changes

a) What is the maximal strength of structural breaks regarding linear judgmental biases so that a linearly corrected judgmental forecast has lower expected error variance than the original forecast?

b) Which error variance reduction can be achieved by extending linear forecast correction methods to consider structural breaks?

Besides the uncertainty resulting from estimating the parameters of the forecast correction model, an additional source of estimation uncertainty exists. As is shown in this work, non-stationarity of time series has to be addressed as it can be an issue for standard forecast correction methods. The different established approaches to ensuring stationarity of time series however differ regarding their properties in terms of the bias–variance trade-off. Most importantly, some approaches require estimating additional parameters, which can in turn increase the variance component of the error. Consequently, RQ 3 aims at identifying which approaches are most beneficial for application in forecast correction.

RQ 3 Forecast Correction – Non-Stationarity
Which approaches to ensuring stationarity of time series provide an additional reduction of the bias component of the error of a linearly corrected judgmental forecast that is higher than the increase of the variance component?

In forecast combination, a key result in the literature, especially on the combination of two forecasts, is the so-called forecast combination puzzle (Stock and Watson, 2004). The puzzle refers to the empirical finding that a simple, unweighted average (SA) of forecasts is typically not outperformed by more complex approaches. Different weighting schemes have different characteristics in terms of the bias–variance trade-off. While the simple average does not necessarily minimize errors on past data, it does not have issues with oversensitivity to the training data. In contrast, more complex methods, such as the so-called optimal weights (OW) proposed by Bates and Granger (1969) that minimize errors in the training data, aim at minimizing the bias component but are highly sensitive to the training data. The trade-off between these two extremes can be addressed by considering the whole spectrum between the two methods by linearly shrinking optimal weights towards the simple average.

Although the performance of the different combination methods is clearly related to the bias–variance trade-off, a decomposition of the combined error variance into a bias and a variance component has not yet been derived in the literature. As the decomposition is an important basis for understanding the performance of different approaches, especially for different shrinkage levels of optimal weights towards the simple average, deriving the necessary theory is of importance for all theoretical analyses of forecast combination.

Given the decomposition, two aspects regarding the bias–variance trade-off are of interest. First, if low shrinkage is used, the sensitivity to the training data is high and a large training sample is required to ensure sufficient stability of the weight estimates. The reverse is true for strong shrinkage. Thus, the required size of the training sample depends on the chosen shrinkage level and can be analyzed to determine how large a sample must be for a combination to be beneficial. Similarly, given a training sample size, a specific degree of sensitivity to the training sample must minimize the expected error variance of the combined forecast. These aspects are formulated in RQ 4 using the common assumption of unbiased errors.

RQ 4 Forecast Combination – Bias–Variance Trade-Off

a) Can the out-of-sample error variance of a combination of unbiased forecasts using optimal weights (OW) shrinked towards the simple average (SA) be decomposed into a bias and a variance component?

b) Given a linear shrinkage level of optimal weights (OW) towards the simple average (SA), which training sample size is required for a combination of unbiased forecasts to have lower out-of-sample error variance than the simple average?

c) Given training sample size, which linear shrinkage level of optimal weights (OW) towards the simple average (SA) minimizes the out-of-sample error variance of a combination of unbiased forecasts?

Similarly to forecast correction, forecast combination is also influenced by structural changes, which result in weights not fitting future observations even if estimated without uncertainty. For a specific combination, small changes up to a certain critical level can be expected not to affect the optimality of the decision. The critical degree of changes can thus be seen as a measure of the robustness of a decision. In general, a more robust decision can be expected to be related to stronger shrinkage since the simple average as an extreme choice are completely independent of structural changes. Thus, the critical changes quantifying the robustness can be used to determine a shrinkage level that is robust against changes up to a certain extent, which can be chosen as a robustness requirement by the expert responsible for the forecast combination. These aspects regarding structural changes and robustness are formulated in RQ 5.

RQ 5 Forecast Combination – Structural Changes

a) How strongly are error covariances allowed to change for a combination of unbiased forecasts to outperform a combination with lower shrinkage of optimal weights (OW) towards the simple average (SA)?

b) Given a robustness requirement in terms of maximum changes of error covariances, what is the optimal linear shrinkage level of optimal weights (OW) towards the simple average (SA) for a combination of unbiased forecasts?

Based on the derived research questions, this thesis makes several contributions to the literature on forecast correction and combination.

Although forecast correction is an established approach to improving the accuracy of judgmental forecasts, this work is the first to analyze the robustness of the approaches against small training samples and structural changes. More precisely, critical values regarding training sample sizes as well as changes of the linear judgmental biases are introduced that result in a corrected forecast having higher accuracy than the original one. Based on the insights into the robustness against structural changes, a new approach is proposed that improves the robustness by explicitly detecting and including structural changes. The approach –as well as the existing approach using exponential weighting of past observations– is empirically evaluated to provide insight into how to address structural changes. Furthermore, non-stationarity is first identified as an issue for forecast correction methods in this work. Several approaches ensuring stationarity of time series are proposed for application in the context of forecast correction and evaluated for different types of time series to provide guidance on which approach to use in specific cases.

In forecast combination, it is well known that the so-called forecast combination puzzle –that complex combination methods do not empirically outperform simple methods– is related to estimation uncertainty and consequently to the bias–variance trade-off. However, this work is the first to aim at decomposing the error of a combined forecast into components related to bias and variance. Based on this decomposition, a minimal training sample is derived for which a combination can be expected to outperform an alternative one, which aids decisions on how to combine forecasts. Furthermore, a new and analytically derived shrinkage level of optimal weights towards the simple average is introduced. This shrinkage level allows explicitly addressing the uncertainty involved in estimating weights from a limited set of past observations. The additionally derived critical changes first provide the means to assess the robustness of a decision for a combination. For this purpose, the changes quantify how strongly error characteristics are allowed to change for a decision for a combination, i.e., a shrinkage level, to still perform better than the robust alternative of using the simple average. Lastly, the robust shrinkage factor, which can be derived from the critical changes, is the first combination approach that explicitly addresses the robustness of a decision.

Overall, this thesis provides a variety of new insights into forecast correction and combination and develops new methods and guidance for successful and robust application of the methods in practical settings.

1.3 Structure of the Thesis

For the purpose of thorough analysis and evaluation of forecast correction and combination, this thesis is structured as follows. The first part focuses on the foundations of the work in this thesis. First, after the introduction and motivation, judgmental forecasting and its strengths and weaknesses in comparison to statistical model-based forecasting methods are discussed in combination with the mechanisms that can be used to integrate judgmental forecasts and statistical models in order to use the strengths of both approaches. As the integration methods require estimating parameters, relevant basic concepts of statistical learning theory are discussed subsequently. Most importantly, the bias–variance trade-off is introduced, which affects the integration mechanisms and is used throughout this work.

The second part analyzes theoretical properties and bias–variance trade-offs in forecast correction and combination methods as integration mechanisms for judgmental forecasts and statistical methods. The robustness against small training samples and structural breaks is assessed and approaches are developed that address the involved bias–variance trade-offs. The proofs of all theorems derived or displayed in this part are provided in the appendix.

The proposed approaches are evaluated in an empirical case study in the third part of this work. The case study in corporate cash flow forecasting and the data from a sample company used for this purpose are introduced. The available forecast data set is then used to evaluate which approaches to solving the bias–variance trade-offs in forecast correction and combination not only are advantageous regarding their theoretical properties but also perform well in practice.

The fourth and final part of this work concludes and provides an outlook on possible extensions and future work.

A graphical overview of this thesis with its four parts is additionally provided in Figure 1.1.

The analyses and evaluation of forecast correction methods are in parts based on Blanc and Setzer (2015b) and Blanc and Ruchser (2016). Furthermore, the theoretical analyses and discussions of forecast combination are partially based on Blanc and Setzer (2016a,b).

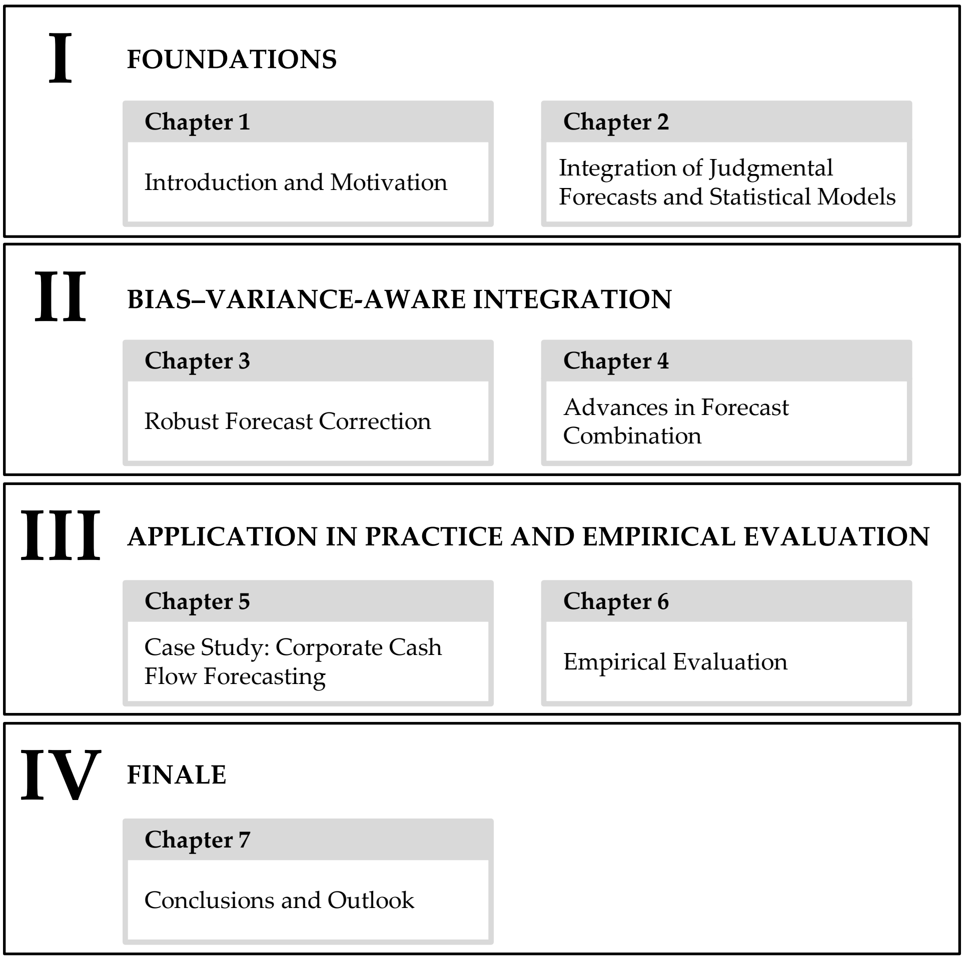

I
FOUNDATIONS
Chapter 1
Introduction and Motivation
Chapter 2
Integration of Judgmental Forecasts and Statistical Models
II
BIAS–VARIANCE-AWARE INTEGRATION
Chapter 3
Robust Forecast Correction
Chapter 4
Advances in Forecast Combination
III
APPLICATION IN PRACTICE AND EMPIRICAL EVALUATION
Chapter 5
Case Study: Corporate Cash Flow Forecasting
Chapter 6
Empirical Evaluation
IV
FINALE
Chapter 7
Conclusions and Outlook

Figure 1.1: The thesis is structured into four parts. The first part provides an introduction and motivation to the thesis and introduces important foundations. The second part focuses on analyzing and improving the mechanisms for integrating judgmental forecasts and statistical models, which are then applied and evaluated in the third part. The last part concludes and gives an outlook on extensions and future work.

Chapter 2

Integration of Judgmental Forecasts and Statistical Models

JUDGMENT, as an established approach to time series forecasting in practice, has a variety of strengths and weaknesses in comparison to alternative model-based forecasts. In the first section of this chapter, a short introduction to judgmental forecasting as well as a discussion of the strengths and weaknesses are provided. Forecast correction and combination as integration mechanisms for judgmental forecasts and statistical models aim at using the strengths of both approaches to increase forecast accuracy. Both methods are introduced in detail in Section 2.2, where a short overview of alternative integration mechanisms is additionally provided. As both forecast correction and combination require estimating a statistical model with parameters, the approaches can be analyzed using results from the statistical learning theory. Relevant basic concepts, especially the bias–variance trade-off, are introduced in Section 2.3 as a foundation for the analyses in this work.

2.1 Judgmental Forecasting

Time series forecasting in general aims at predicting future values of a variable of interest given available information. Important information for a forecast task are on the one hand past observations of the time series as a temporal dependency or development of the variable can often be suspected. On the other hand, available contextual quantitative and qualitative information can be relevant and thus be included in a forecast.

More formally, a forecast F_{t+h} is produced at time t and predicts the value at time $t+h$ where h is the forecast horizon. When a forecast F_{t+h} is produced, the past realized, or actual, values $A_1, \ldots, A_t$ are known and can be used as a basis for the forecast. The predicted variable is realized at time $t+h$ and the actual value A_{t+h} is then known. The accuracy of the forecast can then be assessed, for instance using the mean squared error (MSE) defined in Equation 2.1, which quantifies the accuracy for a set of m forecasts $F_1, \ldots, F_m$ in comparison to the corresponding actual values $A_1, \ldots, A_m$. Since the MSE involves a quadratic penalization of errors, large errors have a much larger influence on the MSE than small errors.

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (A_i - F_i)^2 \tag{2.1}$$

Various statistical time series forecasting models are available, which aim at minimizing the forecast errors. See De Gooijer and Hyndman (2006) for an extensive review of model-based time series forecasting. The model-based approaches provide forecasts with reasonable forecast accuracy and sufficient time series data for estimating the models is often available. As model-based forecasts are cheap to calculate and "particularly simple models can outperform humans in a wide range of situations" (Hogarth and Makridakis, 1981), statistical models could be expected to dominate forecasting tasks in practice.

However, judgmental forecasts are nevertheless often used in practice. For instance Sanders and Manrodt (2003) found that only 11 % of 240 surveyed US companies use software in the forecasting process. See Webby and O'Connor (1996) for an overview of the use of judgment in forecasting. Consequently, judgmental forecasts seem to have strengths different from those of model-based forecasts. For instance Makridakis (1988), Blattberg and Hoch (1990), Webby and O'Connor (1996), and Sanders and Manrodt (2003) surveyed and discussed the strengths and weaknesses of judgmental and model-based forecasting. A key result of the surveys is that the relevance of contextual information is the main reason for the continuing importance of judgment in forecast tasks. Humans producing judgmental forecasts can include all kinds of qualitative information besides the available quantitative information into their forecasts. Statistical forecasting models are in contrast restricted to the limited set of quantitative information that are explicitly included in the model. The qualitative contextual knowledge is especially important if the variables to be predicted are relatively unstable, for instance as a result of changing external factors. Unstable time series are often easier to predict for human experts with contextual knowledge, which allows anticipating changes to a certain extent. These aspects offer a comprehensible explanation of the persistent importance of judgmental forecasting in practice, where contextual factors are likely to be important.

In contrast to the key advantage of being able to include contextual information, judgmental forecasts often exhibit various weaknesses. Most importantly, the judgmental forecasts are often influenced by cognitive heuristics and biases, as for instance reviewed by Hogarth and Makridakis (1981), Bunn and Wright (1991), Goodwin and Wright (1994), and Lawrence et al. (2006).

For example the anchoring and adjustment heuristic in time series forecasting (Hogarth and Makridakis, 1981; Lawrence and O'Connor, 1995) results in humans using the last available value as anchor and then adjusting until a reasonably plausible forecast value is reached. The adjustments from the last value resulting from this procedure are however often insufficient, which can in turn considerably decrease forecast accuracy.

Beyond cognitive heuristics, different time series characteristics are known to negatively influence the accuracy of judgmental forecasts. For instance randomness in time series is often found to influence judgmental forecasts, likely because humans tend to interpret randomness as a signal, as researched by Lopes and Oden (1987), Andreassen (1988), and Harvey (1995), amongst others. Similarly, Reimers and Harvey (2011) found that autocorrelation of time series influences judgmental forecasts. While judgmental forecasts are in general sensitive to autocorrelation, uncorrelated time series are often judged with positive autocorrelation. Furthermore, the autocorrelation of the time series, which was analyzed and predicted before the current one, was found to influence the judgmental forecast.

Overall, judgmental forecasts are likely to be biased and influenced by various factors, which in turn results in reduced forecast accuracy. Biased judgmental forecasting and decision-making are phenomena observed in many contexts and negative effects on business performance can often be found, as discussed by Leitner and Leopold-Wildburger (2011). For instance, in a study by Lawrence et al. (2000), errors in sales forecasts of three manufacturing-based companies were attributed mainly to inefficiencies and biases. In another study, Enns (2002) analyzed the influence of biased and uncertain demand forecasts on production scheduling and found that biases significantly influence lateness of deliveries.

The effects biases and heuristics can, at least partly, be mitigated with well-designed decision support systems, as for instance shown for the anchoring and adjustment heuristic by Remus and Kottemann (1995) and George et al. (2000). However, it is regularly observed that information provided by decision support systems is undervalued by the user, which can result in biases being only partly removed (Lim and O'Connor, 1996; Bhandari et al., 2008).

As biases, which and are likely to exist in judgmental forecasts, influence forecast accuracy and cannot be completely removed by providing decision support to the forecaster, alternative means aiming at improving forecast accuracy are of interest. The integration with statistical models, which is introduced in the next section, provides such methods aiming at improving the accuracy.

2.2 Integration Mechanisms

In general, integration mechanisms for judgment and statistical models aim at improving the accuracy of forecasts by combining the strengths of subjective judgmental forecasts and objective statistical methods in an advantageous way. Goodwin and Wright (1994), Webby and O'Connor (1996), Armstrong and Collopy (1998), Goodwin (2002), Sanders and Ritzman (2004), and Lawrence et al. (2006) provided overviews of the different approaches aiming at improving the accuracy of judgmental forecasts. Four important approaches can be identified from the surveys.

First, in the judgmental adjustment approach, a model-based forecast is produced first. The forecast is then presented to a human expert who can adjust the forecast. The expert is expected to only deviate substantially from the objective model-based forecast in case of strong evidence for an alternative one.

Second, in judgmental bootstrapping, going back to Bowman (1963), the relationship between all available input data and the final judgmental forecast is analyzed. Relevant input data and weightings are derived and a statistical model of the original judgmental forecasts is estimated, which can be used to produce new forecasts.

Third, forecast correction, as introduced by Theil (1966), aims at detecting linear biases in past forecasts. The identified biases can then be removed from new forecasts in order to derive a more accurate corrected forecast.

Fourth, a combination of judgmental and model-based forecasts has been identified as a promising approach as they are likely to have very different strengths and weaknesses, as previously discussed in Section 2.1. In a combination of forecasts, the errors of the different forecasts can cancel each other out, which in turn reduces the error variance.

The first two approaches, adjustments of model-based forecasts and judgmental bootstrapping, are not considered in this work for two reasons. First, forecast correction and combination are more likely to reduce biases in judgmental forecasts as forecast correction directly aims at removing biases and forecast combination reduces biases in a combination with an unbiased model-based forecast. In contrast, biases are likely to persist in the judgmental bootstrapping approach (as the biased judgmental forecasts are analyzed) and biases similar to those in a judgmental forecast are likely to occur in judgmental adjustments of model-based forecasts. Second, judgmental bootstrapping requires a great variety of relevant input data in order to reproduce the judgmental forecast and implementing judgmental adjustment of model-based forecasts requires changing the forecasting process. However, this work is on the one hand not an experimental one. On the other hand, no data set of quantitative and contextual factors relevant to a forecaster is available.

This selection is additionally supported by results in the literature, which indicated that the two excluded approaches often do not improve forecast accuracy. While early studies found considerable evidence in favor of judgmental bootstrapping (Kleinmuntz, 1990), these results were later explained by unrealistic assumptions, such as irrelevance of contextual information and input data without autocorrelation (Bunn and Wright, 1991; Lawrence and O'Connor, 1996). Under more realistic assumptions, the model of the judgmental forecast is found not to outperform the original forecast. Judgmental adjustment of model-based forecasts are designated as "the least effective way to combine statistical and judgmental forecasts" by Sanders and Ritzman (2001) in a review of studies on judgmental adjustments. Model-based forecasts are adjusted very often, too weakly for complex and too strongly for simple models (Franses, 2008). An optimism bias can be suspected for positive adjustments, as these adjustments are less beneficial (Syntetos et al., 2009). Furthermore, small adjustments are likely to decrease accuracy (Fildes et al., 2009) and adjustments are frequently too large in volume (Franses and Legerstee, 2010). These effects may be due to the experts having a different loss functions or due to overconfidence (Franses, 2013).

The forecast correction and combination approaches, which are the focus of this work, are introduced in greater detail next.

2.2.1 Correction of Judgmental Forecasts

The most common approach to forecast correction was introduced by Theil (1966) and is commonly called Theil's method. The approach is based on a decomposition of the mean squared error and is introduced in Theorem 2.1. It should be noted that the following analyses and discussions do not consider the forecast horizon of the judgmental forecasts as Theil's method can be applied for each forecast horizon independently.

Theorem 2.1 (Theil's Decomposition of the MSE). *Given means μ_F, μ_A and standard deviations σ_F, σ_A of the forecasts and actuals, and correlation ρ between forecasts and actuals, the MSE can be decomposed to*

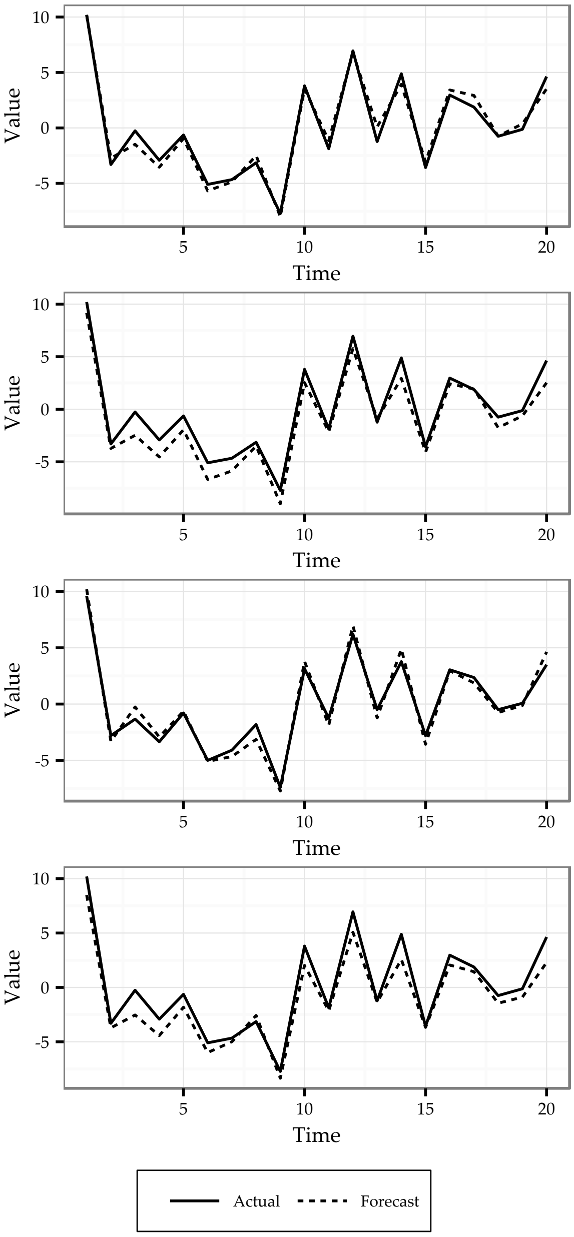
$$MSE = (\mu_A - \mu_F)^2 + (\sigma_F - \rho\sigma_A)^2 + \left(1 - \rho^2\right)\sigma_A^2$$

Interestingly, the terms of Theil's composition have intuitive interpretations. The first term, $(\mu_A - \mu_F)^2$, is the squared difference between the means of the actuals and of the forecasts. The corresponding bias is consequently called *mean bias* and refers to a constant shift of the forecasts in comparison to the actuals.

In contrast, the second term $(\sigma_F - \rho\sigma_A)^2$ compares the standard deviation of the forecasts to the standard deviation of the part of the actuals which is correlated with the forecasts. The corresponding bias is referred to as *regression bias*. Several cases can be imagined regarding the standard deviations and correlation of forecasts and actuals. If forecasts and actuals have equal standard deviation and a high correlation, the deviations from the mean match well and there is no regression bias. In contrast, a bias exists for instance if there is no correlation between forecasts and actuals (the forecasts are random in this case and do not have explanatory power) or if forecasts and actuals are correlated but the standard deviation of the forecasts exceeds the one of the actuals. In the latter case, small values are slightly overestimated whereas high actuals are strongly overestimated. The name regression bias is motivated by the fact that forecasts are (linearly) correlated with the actuals but do not scale perfectly with the actuals.

The last term, $\left(1 - \rho^2\right)\sigma_A^2$, reflects the part of the variance of the actuals that is not included in the forecasts, i.e., the uncorrelated component. The unsystematic error component reflected by this term is consequently called *random error*.

The three types of biases are illustrated in Figure 2.1 where forecast with different biases (dashed lines) are displayed for a short time series of actual values (solid line). Additionally, the values of the three terms of the decomposition are shown in the right part of the figure. In the first example, the forecasts match the actuals well with only random fluctuations around the actual values. Consequently, the MSE consists completely of the random error. In contrast, the second and third example display forecasts with a mean or a regression bias. For the mean bias in the second example, the forecasts are systematically lower than the actual values. In the example with the regression bias, values close to zero have small errors whereas higher (positive or negative) actuals are overestimated. Lastly, the fourth example shows a combination of the two biases, where the actuals are systematically underestimated and the errors scale with the actual values. This last example already resembles real-world time series and forecasts.



MSE	0.475	(100 %)
Mean Bias	0.000	(0 %)
Regression Bias	0.000	(0 %)
Random Error	0.475	(100 %)

MSE	1.475	(100 %)
Mean Bias	1.000	(67.8 %)
Regression Bias	0.000	(0 %)
Random Error	0.475	(32.2 %)

MSE	0.475	(100 %)
Mean Bias	0.000	(0 %)
Regression Bias	0.190	(40 %)
Random Error	0.285	(60 %)

MSE	1.665	(100 %)
Mean Bias	1.000	(60.1 %)
Regression Bias	0.165	(9.9 %)
Random Error	0.500	(30 %)

Figure 2.1: Examples of biases considered by Theil's method. The forecasts in the first example are unbiased and fluctuate randomly around the actuals. In contrast, the forecasts in the second and third example have a mean and a regression bias respectively. Both biases are present in the third example.

Based upon the decomposition of the MSE, Theil (1966) proposed the optimal linear correction of forecasts introduced in Theorem 2.2, which removes the two systematic components of the error, i.e., the mean and the regression bias.

Theorem 2.2 (Optimal Linear Correction)**.** *Mean and regression bias in Theil's decomposition can be eliminated by calculating a corrected forecast*

$$F_C = \beta_0 + \beta_1 F$$

where β_0, β_1 *are the regression coefficients in the linear regression* $A = \beta_0 + \beta_1 F + \epsilon$*.*

The correction of the biases can also be derived from a forecast-actual plot, which is shown in Figure 2.2 for the previous example. In case of an unbiased forecast, the forecasts are randomly distributed around the line where the forecasts are equal to the actuals, which is indicated by a dashed line in the plot. The upper left example clearly indicates unbiasedness and the regression slope is one while the intercept is zero. In contrast, the intercept is unequal zero in the upper right figure (mean bias) whereas the slope is unequal one in the lower left figure (regression bias). If mean bias as well as regression bias exist in a forecast (lower right figure), the intercept is unequal zero and the slope is unequal one. In each case, the regression coefficients indicate how the forecasts have to be transformed in order to match the actuals in expectation.

Empirical case studies applying Theil's method in different domains have shown that forecast correction often increases forecast accuracy substantially.

In an early study, Moriarty (1985) corrected one single sales time series spanning six years. The author found mean biases in four of the six years and additional significant regression biases. A correction of the biases led to a significant reduction of the mean squared error for two of the years. In contrast to the years for which the correction was advantageous, the other two years primarily exhibited a much lower mean bias.

Bohara et al. (1987) used Theil's method to correct macroeconomic forecasts (GNP, real GNP, GDP deflator, and corporate profits after taxes) from the ASA-NBER survey, which were the result of a survey amongst 50-60 forecasters. In the empirical evaluation, the correction substantially increased the out-of-sample errors for all four time series. It should however be noted that the ASA-NBER forecasts are already combinations of multiple forecasts and consequently likely to be less biased than individual forecasts and that all four time series are not stationary and show clear trends.

llmakunnas (1990) analyzed forecast correction methods from a pretesting point of view, i.e., assuming that a statistical test on past performance is used

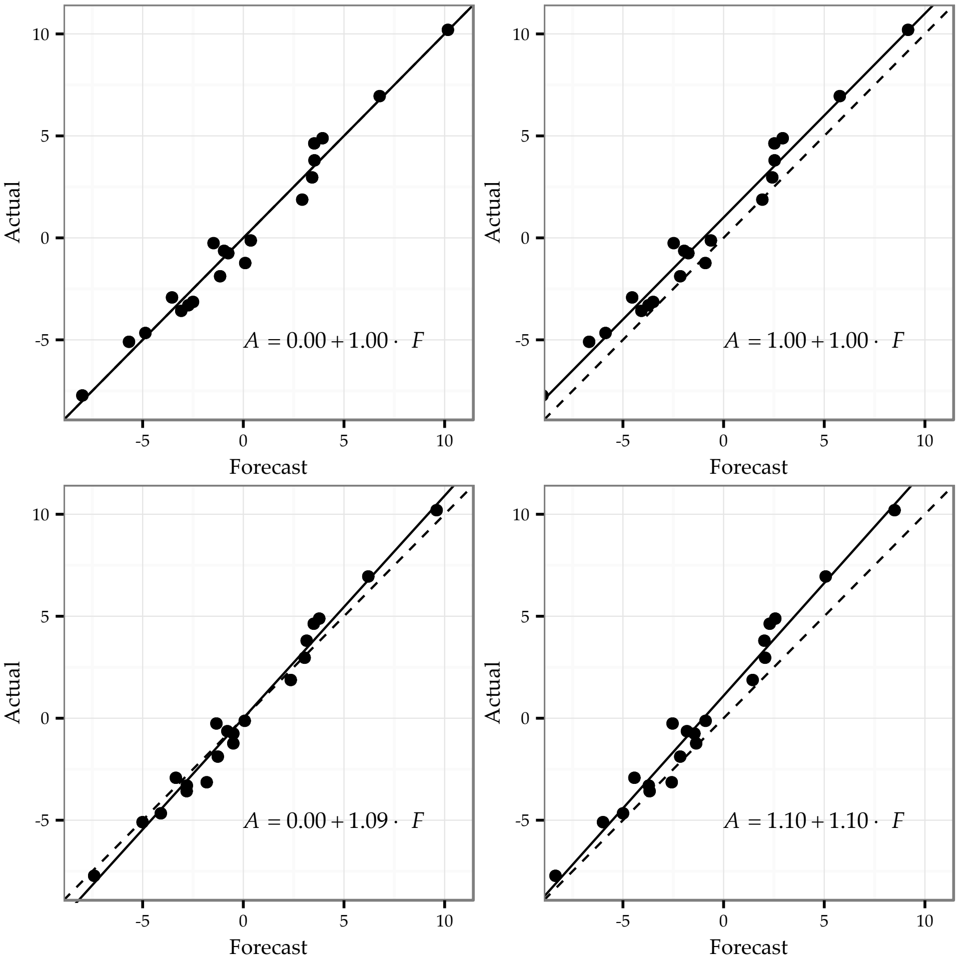


Figure 2.2: In a forecast-actual plot, the different biases and combination of biases correspond to different patterns. In the unbiased case (upper left), the points are randomly distributed around the dashed line where the forecasts are equal to the actuals. In contrast, in case of a mean (regression) bias in the upper right (lower left) figure, the regression line through the points is shifted (tilted) in comparison to the unbiased line. If both biases exist (lower right figure), the regression line is shifted and tilted.

to decide whether the original forecast, a naïve forecast or the corrected forecast is used. Based upon the results, the author cautioned against using forecast correction and suggested to use forecast correction only if the correction is superior at a high confidence level or if systematic and persistent biases are likely to occur.

Elgers et al. (1995) corrected 6,302 yearly earnings forecasts generated by analysts. The mean squared error of the forecasts could be reduced significantly.

Extensive experiments were conducted by Goodwin (1996, 2000), both in a laboratory setting and on empirical data. In the laboratory experiment, students produced one-period-ahead forecasts, for which a correction using Theil's method was shown to increase accuracy. Special periods in the form of promotional events were included in order to check for robustness against sporadic events. The empirical experiments were based on forecasts from three companies from different industries covering up to 42 months of sales data. The corrected forecasts outperformed the original expert forecasts for two of the companies. However, the improvements were only significant for one of the companies under study. Additionally, the business impact of forecast correction was evaluated using a loss function with asymmetrical penalties for under- and overestimation. It was shown that using corrected forecasts could have reduced costs resulting from under- or overproduction as a result of inaccurate forecasts by 26-54 %, depending on the assumed ratio between costs of under- and overproduction.

Goodwin (1997) used Theil's method with discounted weights for older observations for the correction of forecasts resulting from a laboratory experiment. The author found that using discounted weights of past errors led to lower mean absolute errors for several types of time series. Significant improvements over Theil's method were observed for low-noise time series. For high-noise time series, both approaches showed comparable performance, except for time series with a trend reversal, where the weighted approach performed better, and a time series with a linear trend, where Theil's method led to lower errors.

In an exploratory study, Shaffer (1998) analyzed 33 one-quarter-ahead forecasts of a macroeconomic indicator produced by one expert and also found the forecasts to be biased. As a result of the correction, the mean squared error was reduced by nearly 25 %.

In a case study on demand forecasts of a cell phone company, Utley (2011) showed that using an alternative, more robust, estimation procedure can improve the performance of Theil's method. Robust regression methods were also applied successfully for detecting (and predicting) errors in analysts' earnings forecasts (Boudt et al., 2014).

Blanc and Setzer (2015b) compared forecast correction methods in a case study on corporate cash flow forecasts and found that exponential weighting is required to achieve significant and robust accuracy improvements.

Theil's decomposition and correction was also applied for analyses of the forecasting skill and the correction of model-based forecasts, see for instance Brandon et al. (1983), Ashley (1985), and Stewart and Reagan-Cirincione (1991).

Overall, forecast correction is an established approach that is based upon a solid theoretical foundation in the form of Theil's decomposition. However, as Theil's method focuses on two specific linear biases, only these biases can be identified and removed. Other biases can only be addressed implicitly and no additional information is added by the correction. The combination of forecasts, which is introduced next, likewise allows implicitly addressing various biases.

2.2.2 Combination of Judgmental and Model-Based Forecasts

For a reduction of biases and errors independently of specific linear biases, a combination of the judgmental with a model-based forecast can be used. As a combined forecast should be within the interval spanned by the judgmental forecast and the unbiased model-based forecast, all systematic biases in the judgmental forecast are likely to be reduced. In general, statistical approaches to time series forecasting aim at detecting patterns, such as trend or seasonality, in past values. If systematic patterns are detected, the patterns can be extrapolated and to derive forecasts for future values of the variable of interest. For example De Gooijer and Hyndman (2006) and Makridakis et al. (2008) provide extensive reviews of approaches to time series analysis and forecasting.

Empirical research indicates that autoregressive integrated moving average (ARIMA) models often represent patterns in time series well and often result in good predictive accuracy (De Gooijer and Hyndman, 2006; Cryer and Chan, 2008). The main reason for this is that ARIMA models are very flexible and different models and parameterizations can be used for various kinds of time series. Damped trend exponential smoothing, another method that is often recommended (Gardner, 2006), is less flexible in comparison to ARIMA and always uses the local trend as a basis. The trend is however dampened in order to "introduce a note of conservatism" (Gardner and McKenzie, 2011).

As an alternative to choosing either the statistical or the judgmental forecast (or between different judgmental or model-based forecasts), a combination of forecasts can be derived. Forecast combination is used in economics since the work of Reid (1968) and Bates and Granger (1969) and has been shown to increase forecast accuracy in many scenarios, as surveyed by Clemen and Winkler (1986) and Timmermann (2006), amongst others. The combination of forecasts and its effectiveness is motivated by the simple idea of a portfolio diversification effect, as already argued by Bates and Granger (1969). In a combination of forecasts with differing error characteristics, errors cancel each other out at least partially. Furthermore, extreme errors of one forecast can be compensated by the other forecasts included in the combination, overall resulting in a substantial reduction of the error variance of the combined forecast. Even a combination of misspecified forecasting models can outperform individual well-specified forecasts (Hendry and Clements, 2004).

Technically, given a vector of k forecasts $\vec{F} = (F_1, \ldots, F_k)$, forecast combination is a linear combination of the forecasts defined as $F = w^\top \vec{F}$ with weights $w \in \mathbb{R}^k$

and $\sum w_i = 1$. The weights w can be defined in advance or learned from past forecast errors. In the latter case, the weights are estimated from one set of forecasts and corresponding errors (the training data) and then applied to another, previously unknown set of forecasts (the evaluation data). The aim of forecast combination is to choose weights w that perform well on the evaluation data, even though only a limited set of training data is available.

A simple approach where weights are defined in advance is the simple average (SA), which weights the k forecasts equally. SA corresponds to the weight vector w^S in Equation 2.2, where $\vec{1}_k$ is a vector of ones of length k.

$$w^S = \frac{1}{k}\vec{1}_k \tag{2.2}$$

The so-called optimal weights (OW), which were introduced by Bates and Granger (1969) for the basic case with two forecasts, can be considered as the complement to SA. While SA does not take available training data into account, OW minimizes the combined error variance within the training sample. The multivariate generalization of the optimal weight estimate $\hat{w}^O$ (see for instance Timmermann (2006)), which uses an estimate of the error covariance matrix in the training sample, $\hat{\Sigma}$, is introduced in Equation 2.3.

$$\hat{w}^O = \frac{\hat{\Sigma}^{-1}\vec{1}_k}{\vec{1}_k^\top \hat{\Sigma}^{-1}\vec{1}_k} \tag{2.3}$$

The two approaches can be used to illustrate the error variance reduction effect of forecast combination. The combined error variance on the training data of a combination of two forecasts with SA and OW is shown in Figure 2.3. The error variances of the two individual forecasts $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$ are indicated by dotted, horizontal lines in the figure.

The in-sample error variance of OW never exceeds the lowest error variance amongst the individual forecasts. Error variances with SA are lowest for error correlation -1, and then monotonically increase, at some level of error correlation exceeding the variance of the best individual forecast, but never exceeding the error variance of the worst individual forecast. The difference between error variances with SA and OW is small for strong negative correlations and largest for strong positive correlations.

Although the basic variance reduction effect is established, an important question discussed in the literature is how many (and which) forecasts to include in a combination to ensure the diversification effect and the resulting reduction in error variance. Armstrong (2001) recommended to include at least five forecasts.

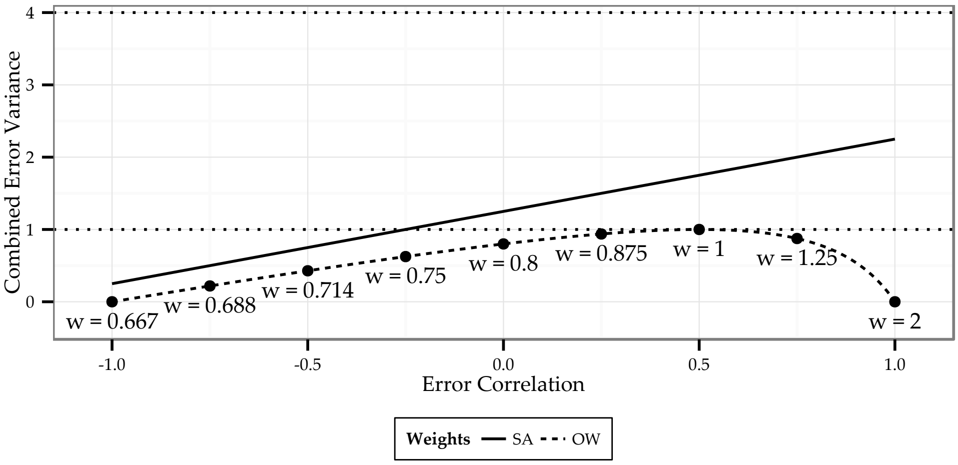


Figure 2.3: Combined in-sample error variances for SA and OW combination of two forecasts with error variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$, depending on the error correlation. With SA, the combined error variance is always lower than the error variance of the worse forecast and can be lower than the error variance of the better forecast, depending on the error correlation. For OW, weights depend on the error correlation and are displayed in the graph. The resulting combined in-sample error variance are –in the worst case– equal to the error variance of the better forecast.

Davis-Stober et al. (2014) suggested constructing an optimally wise group by selecting judgments that are as negatively correlated with each other as possible and to even include less informed or biased forecasts to ensure diversity. Some research suggests that selecting a subset of forecasts dominates averaging over too many forecasts. For instance Mannes et al. (2014) suggested a select-crowd strategy, which ranks judges by accuracy and only includes the opinions of the top judges. Similarly, Budescu and Chen (2015) improved the aggregated judgment by eliminating poorly performing individuals from the crowd. The number and choice of forecasts included in a combination influences the combined error variance in two ways. On the one hand, an increasing number of forecasts allows finding weights that result in a large portion of the errors of the forecasts canceling each other out. On the other hand, including more forecasts requires estimating additional weights, which can increase errors because of estimation errors. The consensus of including a number of forecasts that ensures diversity while not including too many forecasts consequently tries to find a reasonable trade-off between these two aspects.

Combining one or several judgmental and model-based forecasts to derive a forecast that uses the strengths of both judgment and statistical models is a special case of the general forecast combination. Although most of the literature focuses on forecast combination in general or on the combination of different model-based forecasts, promising results on the combination of judgmental and model-based forecasts have been found in empirical studies.

Lawrence et al. (1986) averaged judgmental forecasts and deseasonalized single exponential smoothing forecasts of 68 monthly time series from the M Competition (Makridakis et al., 1982). The experiments showed that the benefits from combining judgmental and model-based forecasts exceed those of combining only model-based forecasts. The combination is most beneficial for short forecast horizons and time series that are easier to predict.

Bohara et al. (1987) combined ARIMA forecasts of four macroeconomic time series (GNP, real GNP, GDP deflator and corporate profits after taxes) with forecasts from the ASA-NBER survey using a simple average as well as unconstrained and constrained (zero intercept and weights summing up to unity) multivariate regression. The simple average outperformed both individual forecasts out-of-sample for two of the four time series. For the GNP deflator, the simple average combination outperformed the ASA-NBER survey forecast. The unrestricted combination as a more flexible approach performed worse than the simple average for all time series. Similarly, the restricted combination only outperformed the simple average for the GNP deflator time series. It should however be noted that the ASA-NBER forecasts are already combinations of multiple judgmental forecasts.

Blattberg and Hoch (1990) used equal weighting of judgmental and custom-made model-based forecasts of catalog fashion sales in two companies and coupon redemption rates in three companies. The variance explained by the combined forecast on average improved 16 % over the best individual forecast.

Lobo and Nair (1990) analyzed combinations of analyst forecasts with corresponding statistical forecasts of earnings of 96 companies for the years 1976 to 1983. As judgmental forecasts, a forecast from the Value Line Investment Survey and an average of analyst forecasts from the Institutional Brokers Estimate System were used. Statistical forecasts were produced using an ARIMA model and a random walk with drift model, each calibrated using 60 quarters of data. The authors found that combinations outperformed all individual forecasts. Furthermore, combinations using cross-sectional weights learned from two years of past data outperformed equal weights of forecasts. Combining a higher number of forecasts further increased accuracy.

Sanders and Ritzman (1990) combined daily demand forecasts of 22 time series covering three years. Forecasts were produced by warehouse planners and by different statistical models. It was found that the model-based forecasts already outperformed the judgmental forecasts, while the combination of judgmental and model-based forecast improved accuracy even further. In an ensuing experiment, Sanders and Ritzman (1995) compared forecasts with contextual knowledge (produced by warehouse managers with information from real-world experience) to forecasts with technical knowledge (produced by students with knowledge on forecasting models and data analysis). The authors found that combinations including the forecasts with contextual knowledge significantly dominated other combinations, indicating that contextual knowledge is essential for a contribution to the combined forecast.

Goodwin (2000) combined judgmental forecasts of eight time series of quarterly sales figures (produced by 16 students with technical knowledge) with model-based forecasts calculated by the Forecast Pro software. Furthermore, judgmental sales forecasts from two companies were combined with exponential smoothing forecasts. In the experiment with students, the combined forecasts improved over the judgmental forecasts but did not outperform the model-based forecasts, except for special periods, where the students had additional information. In contrast, the combination of the real-world judgmental forecasts resulted in improvements over all individual forecasts for both companies.

Franses (2011) analyzed why averaging of statistical and judgmental forecasts works although the expert forecasts are regularly found to be biased. The analysis was based upon the idea that a judgmental forecast can be decomposed into a reproducible part (which can be modeled using bootstrapping of judgmental forecasts) and a non-reproducible part (the intuition). In this setting, the reproducible part makes the combination work whereas the intuition part is of smaller importance and is partly averaged out.

Overall, considerable evidence exists that forecast combination is not only advantageous in general but that the combination of judgmental and model-based forecasts is often especially beneficial for the forecast accuracy. While simple averaging of forecasts is often successfully applied, alternative combination methods, such as optimal weights, exist that involve estimating parameters, i.e., the weights of the included forecasts.

2.3 Statistical Learning Theory

As weights have to be estimated from available data in forecast correction as well as forecast combination, the statistical learning theory aiming at understanding statistical model involving estimated parameters can be applied. In this section, basic aspects of the statistical learning theory, especially the so-called bias–variance trade-off, are introduced based upon Hastie et al. (2009) and James et al. (2014).

In a first step, it can be assumed that there is a functional relationship f between a set of predictors X and the dependent variable (corresponding to the actuals in the forecasting setting) A as shown in Equation 2.4, where ϵ is a random error term with mean zero that is independent of X.

$$A = f(X) + \epsilon \tag{2.4}$$

If the true functional relationship is known, the values of the predictors can be plugged into the function, resulting in a prediction or forecast $F = f(X)$ with an error only consisting of the random error ϵ. However, it is rarely the case that the true function relationship is known, except for instance for physical laws. Consequently, the function f has to be estimated as $\hat{f}$ from available observations. Using the estimated function, a prediction can be calculated as $F = \hat{f}(X)$.

In contrast to the case with known f, the prediction error not only depends on the random error ϵ if the functional relationship is estimated. If the estimate $\hat{f}$ is very different from the true function f, errors are likely to be substantially higher than for a good estimate $\hat{f}$. Thus, the accuracy of the estimate of f can also influence the prediction error.

As for instance shown by James et al. (2014), the expected MSE can be decomposed into parts that are related to properties of $\hat{f}$, as shown in Equation 2.5.

$$\mathrm{E}\left[\left(A - \hat{f}(X)\right)^2\right] = \left(\mathrm{Bias}\left[\hat{f}(X)\right]\right)^2 + \mathrm{Var}\left[\hat{f}(X)\right] + \mathrm{Var}\left[\epsilon\right] \tag{2.5}$$

In this decomposition, the bias and the variance of $\hat{f}(X)$ are defined as shown in Equations 2.6 and 2.7. Throughout this work, the terms bias component and variance component are used when referring to the bias and variance parts of the expected MSE in order to avoid confusion with the general statistical variance and with judgmental biases.

$$\mathrm{Bias}\left[\hat{f}(X)\right] = \mathrm{E}\left[\hat{f}(X)\right] - f(X) \tag{2.6}$$

$$\mathrm{Var}\left[\hat{f}(x)\right] = \mathrm{E}\left[\left(\hat{f}(X) - \mathrm{E}\left[\hat{f}(X)\right]\right)^2\right] \tag{2.7}$$

It should be noted that the expectation and variance in the decomposition result from applying different $\hat{f}$ estimated from different samples from the sample population to all possible X. The bias component then reflects how much the values $\hat{f}(X)$ in expectation deviate from the true values $f(X)$. If the expectation of different estimates of f equals the true function, there is no systematic (constant) bias component of the estimates. A bias component is in general caused by the estimate $\hat{f}$ missing a relevant relationship between X and A. In contrast, the variance component reflects how strong the predictions $\hat{f}(X)$ differ across different estimates. If different estimates lead to very different (similar) results, the variance component is high (low). A high variance component is usually caused by a high sensitivity to different training samples where $\hat{f}$ changes substantially upon random fluctuations in the sample that is used for the estimation.

The last term of the decomposition in Equation 2.5, Var $[\epsilon]$ results from the random error and does not depend on $\hat{f}$. This term is consequently irreducible and imposes a lower bound on the expected MSE. Thus, in order to minimize the expected MSE, the bias and the variance components have to minimized. Unfortunately, bias and variance components cannot be minimized independently as a result of the nature of the two components. The bias component is reduced if all potentially relevant relationships between X and A are included in the estimate $\hat{f}$. To achieve this goal, a high sensitivity is required since all minor fluctuation in the training sample can indicate a relevant relationship that must be included in $\hat{f}$ in order to minimize the bias. However, considering increasingly minor fluctuations in the training sample by definition increases the variance component. Bias and variance components consequently have diametrically opposed behavior when the sensitivity of $\hat{f}$ is changed and the two components cannot be minimized simultaneously. This relationship is illustrated qualitatively in Figure 2.4.

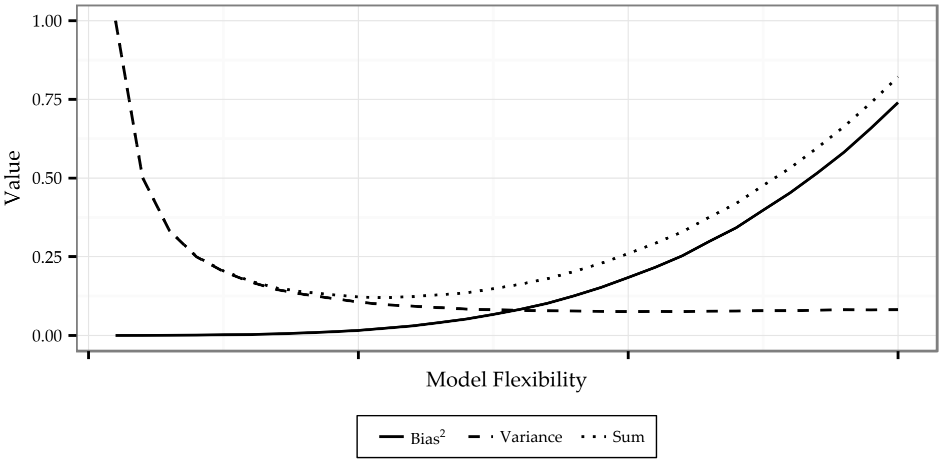


Figure 2.4: Qualitative illustration of the bias–variance trade-off. The squared bias component decreases with increasing model flexibility whereas the variance increases. The MSE, composed of the squared bias component and the variance component, is minimized for a medium model flexibility.

The bias component clearly decreases with increasing model flexibility or complexity whereas the variance component continuously increases. Consequently, there is a certain degree of model flexibility where the sum of the squared bias component and the variance component is minimal, which in turn minimizes the expected MSE.

While the basic relationship between the bias and variance components is clear and well understood, the values of the components given model flexibility are in general unknown. As a consequence, the optimal degree of model flexibility cannot be determined in a straightforward way. For this reason, deriving insights on how the bias and variance components are shaped and how they relate to each other in forecast correction and combination is the issue this work focuses on. Understanding the involved bias–variance trade-offs gives an important orientation on how to successfully and robustly correct and combine judgmental forecasts.

Part II

Bias–Variance Aware Integration

Chapter 3

Robust Forecast Correction

As a variety of biases can influence the accuracy of judgmental forecasts, the correction of judgmental forecasts using Theil's method (Theil, 1966) aims at identifying and removing these biases to improve accuracy. While Theil's method is in theory straightforward and has a solid foundation in the form of the decomposition of the MSE, the robust application in practice is more complex.

The population variances of the forecasts and actuals as well as their correlation, which are used in the decomposition and the optimal linear correction, are unknown in practice. As a consequence, these parameters have to be estimated from available data to identify systematic biases that can then be removed from future forecasts. Section 3.1 gives an overview over the different approaches for estimating the parameters of the correction model that are used in the literature.

As a result of using estimated parameters, the uncertainty caused by the estimation itself as well as by structural breaks influences the accuracy of the corrected forecasts. This aspect is analyzed in Section 3.2. First, as a result of the required parameter estimation and the resulting uncertainty, the statistical learning theory suggests that the error of the corrected forecast has an additional variance component in comparison to the original forecast. This variance component is influenced by the size of the available training sample that is analyzed in a first step. Second, the optimal correction assumes that the biases in future forecasts are equal to those in past forecasts. However, structural changes in forecast errors as a result of biases changing over time can result in estimated biases being inaccurate, which in turn increases the bias component of the error of the corrected forecast. Changes can for instance occur because of learning effects or staff turnover. Consequently, the relevance of structural changes and the impact of breaks is assessed. Third, non-stationarity of time series, which is likely to occur in practice, is proven to be an issue that can severely decrease the performance of Theil's method. While non-stationarity can be addressed, additional parameters have to be estimated for some approaches, which can in turn additionally increase the variance component of the error of the corrected forecast.

The theoretical results for the three aspects are illustrated and discussed in Section 3.3. Based upon the derived insights, Section 3.4 proposes extensions to Theil's method that aim at transferring the theoretical results into practice. The extensions enable addressing structural changes in a more flexible way by detecting and incorporating potential breakpoints while still allowing for discounted weights of old observations. Furthermore, the extensions are designed to enable addressing non-stationarity in different ways.

Finally, Section 3.5 concludes and discusses the implications and the limitations of the analyses and results.

3.1 Model Estimation in Forecast Correction

Theil's decomposition of the MSE and the resulting linear correction, as introduced in Theorems 2.1 and 2.2, use the population means and variances of the forecasts and the actuals as well as their population correlation. However, these population parameters cannot be expected to be known in practice. Thus the coefficients of the linear regression underlying the optimal correction have to be estimated from past data. For this purpose, slope and intercept are estimated as $\hat{\beta}_0, \hat{\beta}_1$ in a regression $A = \beta_0 + \beta_1 F + \epsilon$ with past actuals and forecasts A, F and residuals ϵ.

In the literature, different approaches to estimating the regression coefficients β_0 and β_1 have been used for the application in forecast correction. In general, the coefficient estimation problem can be stated as a minimization of the sum of the penalized errors, as shown in Equation 3.1. The errors are penalized using a function $p : \mathbb{R} \rightarrow \mathbb{R}_0^+$, which is symmetrical, non-decreasing and has unique minimum at zero.

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \sum_t p\left(A_t - \left(\hat{\beta}_0 + \hat{\beta}_1 F_t\right)\right) \tag{3.1}$$

In Theil's method (Theil, 1966), ordinary least squares (OLS) penalties are used, which corresponds to the penalty function presented in Equation 3.2

$$p_{OLS}(\epsilon) = \epsilon^2 \tag{3.2}$$

The OLS estimator results in an estimate that is unbiased and minimizes the variance of the residuals, which are in general preferable properties over alternative estimators. However, an issue well known to influence parameter estimates especially in OLS linear regression are outliers. For instance Cook (1977) and Chatterjee and Hadi (1986) analyzed and discussed the influence of outliers on

OLS estimates of linear regression coefficients. In the case of forecast correction, observations can exist in the training data where the actual value is strongly increased or unexpectedly low, for instance because of unexpected one-time effects or business-related changes, whereas the forecast value is in the normal range of forecast values. Depending on the strength of the effect, these observations can be outliers in the linear regression that strongly influence the regression coefficient estimates.

In linear regression theory, a considerable body of literature focuses on robust regression, i.e., estimators that are to a certain extent robust against outliers or other violations of the assumptions of linear regression. See Rousseeuw and Leroy (2005) for an overview of robust methods in linear regression. A simple approach to robust regression is using least absolute deviation (LAD) estimates instead of OLS estimates. LAD increases the robustness against outliers by using a minimization of the sum of absolute values of the errors instead of squared errors, which reduces the influence of extreme errors. The corresponding penalty function is shown in Equation 3.3.

$$p_{LAD}\left(\epsilon\right) = |\epsilon| \tag{3.3}$$

Ensuring robustness against outliers can also be seen in the context of the bias–variance trade-off. Treating outliers always requires reducing the weight of some observations that are likely to be outliers. As a consequence, the expected fit of the estimated model is worse than for a standard OLS regression without treatment of outliers. The bias component of the error consequently increases, depending on the desired degree of robustness against outliers. However, the variance component decreases since outliers are less likely to influence estimates, resulting in a decreased variance of the estimates.

For application in forecast correction, LAD was proposed by Utley (2011). Using LAD in a linear regression results in robustness against outliers in the independent variable, i.e., against unusually high or low actual values in forecast correction. However, robustness against leverage points, i.e., unusually high or low forecasts, is not achieved and the estimated regression coefficients can still be influenced by outliers of this type (Rousseeuw and Leroy, 2005; Giloni et al., 2006).

As a side-effect, LAD may better reflect a non-quadratic loss function of forecasters, which has been shown to be likely for instance for financial analysts (Basu and Markov, 2004).

Besides outliers, structural changes in the forecast-actual relationship can influence the performance of forecast correction. Technically, a structural change results in the characteristics of the training data differing significantly from those of the data the estimated model is applied to. In terms of the bias–variance trade-off, a structural change results in an increased bias component of the error of the corrected forecast while the variance component is unchanged since parameter estimates are still used. Overall, structural changes can result in substantially increased errors of the corrected forecast and even negate the benefits of correcting judgmental forecasts.

As a consequence, Goodwin (1997) extended Theil's method with discounted weights to give more weight to recent observations. Technically, the approach corresponds to a weighted least squares regression (WLS) with exponential weights if the same discount factor is used for β_0 and β_1. The errors in the minimization problem are weighted geometrically with descending weights γ^t for older observations. The decrease of the weights over time can be controlled by the discount factor $\gamma \geq 1$. The corresponding minimization problem is defined in Equation 3.4, where the weights γ^t are included in comparison to Equation 3.1. While Goodwin (1997) used OLS estimation, the weighted approach can also be used with LAD and the definition consequently uses a generic penalty function p.

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\arg\min} \sum_t \gamma^t p\left(A_t - \left(\hat{\beta}_0 + \hat{\beta}_1 F_t\right)\right) \tag{3.4}$$

Theil's method as well as its variants proposed by Goodwin (1997) and Utley (2011) are obviously special cases of Equation 3.4. All approaches can consequently be treated as special cases of the weighted approach.

The discount factor γ used in the weighted approach introduces an additional parameter which has to be defined in advance or learned from available data. Goodwin (1997) proposed determining the discount factor in a rolling pseudo out-of-sample evaluation for the last 12 data points of the available training data. In this evaluation, a correction model is estimated for different discount factors using Equation 3.4 and OLS and all observations except the last 12 observations. The models are then applied to the first of the twelve hold-out data points. The data point is then added to the training data and the procedure is repeated for the remaining data points. The errors resulting from the correction models with different discount factors are aggregated, for instance using the MSE, and the discount factor with the lowest aggregated error is chosen.

This procedure can be expected to address structural breaks in the training data (for instance a changing bias) in an indirect and implicit way. If a structural break exists in the training data, not only the final correction model but also the pseudo out-of-sample models and their outcomes are likely to be influenced by the structural break. If a structural break exists, weights are chosen that minimize the influence of outdated observations before the structural break while, at the same time, preferring weights resulting in stable models.

Overall, the weighted method with the pseudo out-of-sample identification of the discount factor tries to address the bias–variance trade-off resulting from structural changes in the training data. Treating structural breaks using the approach reduces the bias component and increases the variance component because of the additional parameter and the unevenly distributed weights that result in a lower stability of the estimates. However, if no structural break exists, the bias component is unchanged while the variance component slightly increases. The assumption of the procedure is consequently that the reduction of the bias component in cases with structural changes outweighs the increase of the variance component in these as well as all other cases.

In summary, outliers and structural breaks in Theil's method have been identified as issues and estimation methods that are expected to address these aspects have been proposed for application in forecast correction. Outliers and their effect on linear regression have received much attention in the statistical literature (Rousseeuw and Leroy, 2005). In contrast, while there is awareness that structural changes are an issue, it is currently unknown which effect they have on forecast correction models. More precisely, the means required to assess the robustness of forecast correction against structural changes, i.e., how strongly biases are allowed to change, do not exist. Furthermore, although estimation is always used when applying Theil's method, it is currently unknown how strongly the estimation influences the error of the corrected forecast. This includes the parameters estimation that is additionally required in the preprocessing of the forecasts and actuals before application of forecast correction models. Thus, the next sections focus on understanding the influences of parameter estimation and structural changes on forecast correction. For the theoretical analyses, an unweighted OLS estimation is assumed as the estimator and its properties are, in comparison, well studied.

3.2 Robustness in Forecast Correction

The different estimation procedures introduced and discussed in the previous section all rely on available past data. Thus, a key aspect driving the bias–variance trade-off is the size of the available data, which is analyzed first in this section in order to determine the robustness against small training samples.

Furthermore, structural changes can result in substantial differences between past and future data, which can in turn introduce an additional bias component and increase errors. The robustness of forecast correction methods against structural changes is thus analyzed subsequently. Lastly, non-stationarity is shown to be an issue for forecast correction methods. Although non-stationarity can be addressed, the approaches aiming at ensuring stationarity additionally influence the bias–variance trade-off.

3.2.1 Training Sample Size

In forecast correction models, the parameters are in general estimated from an available set of training data (consisting of past forecasts and realizations), as discussed in the previous section. As the instability of the estimated parameters increases with decreasing training sample size, the variance component can be expected to increase with decreasing number of observations. A relevant question is consequently whether the training sample sizes available in practice are large enough to allow an application of forecast correction methods to be robust and beneficial.

The MSE is a random variable depending on the population characteristics of the samples a correction model is estimated from and applied to. Thus, deriving the expected MSE of the corrected forecast when the parameters are estimated is required for an analysis of the influence of the sample size. Theorem 3.1 presents the MSE of the corrected forecast when the parameters of the correction model are estimated from a training sample of finite size and then applied to a second sample with identical population characteristics.

Theorem 3.1 (MSE of Corrected Forecast With Estimation)**.** *Assuming that the parameter β_1 of the forecast correction model is estimated as $\hat{\beta}_1$ from a finite bivariate sample and applied to a second independent bivariate sample (both with variances of actuals and forecasts σ_A^2, σ_F^2 and correlation ρ). Then the expected MSE is*

$$MSE = \sigma_A^2 + \sigma_F^2 \left(Var\left[\hat{\beta}_1\right] + \left(E\left[\hat{\beta}_1\right]\right)^2 \right) - 2E\left[\hat{\beta}_1\right] \rho \sigma_A \sigma_F$$

Besides the characteristics of the forecasts and errors, the error variance of the corrected forecast (which is equal to the MSE) presented in Theorem 3.1 depends on the expectation and variance of the regression coefficient estimate $\hat{\beta}_1$. Thus, a more detailed analysis of the expectation of the MSE requires a formally defined sampling distribution of the estimate $\hat{\beta}_1$. This distribution is only known under the assumption of bivariate normality of the forecasts and actuals. Using this assumption, the error distribution of the corrected forecast is displayed in Theorem 3.2 for the special case.

Theorem 3.2 (MSE of Corrected Forecast With Estimation Under Normality)**.** *Assuming bivariate normality of forecasts and actuals (with variances of actuals and forecasts σ_A^2, σ_F^2 and correlation ρ), the expected MSE of the corrected forecast is*

$$MSE = \left(1 + \frac{1}{n-3}\right)\left(1 - \rho^2\right)\sigma_A^2$$

As a clear result of Theorem 3.2, estimating $\hat{\beta}_1$ from a training data set increases the MSE of the corrected forecast in comparison to the result without estimation. This result is intuitive and in line with the statistical learning theory. Although the result was obtained under the assumption of bivariate normality of forecasts and actuals, this basic relationship can be expected to be independent of the distribution of the data while the strength of the effect might differ.

Since a finite training sample is used to estimate the coefficients of the forecast correction model, the basic property that the corrected forecast is at least as accurate as the original forecast is no longer necessarily true. Cases with very small training samples can be imagined, where estimation errors result in a very high variance component of the error that negates all potential accuracy increases of the correction. Consequently, a minimal training sample size in many cases exists for which the corrected forecast can be expected to, in expectation, outperform the original (biased) forecast. Again using the assumption of bivariate normality of actuals and forecasts, Theorem 3.3 introduces this minimal sample size.

Theorem 3.3 (Minimal Training Sample Size Under Normality)**.** *Assuming bivariate normality of actuals and forecasts (with variances of actuals and forecasts σ_A^2, σ_F^2 and correlation ρ). Then the minimal sample size, which is required so that the corrected forecast can be expected to have a lower MSE than the original biased forecast, can be derived as*

$$\mathring{n} = \lceil \frac{\left(1-\rho^2\right)\sigma_A^2}{\left(\mu_A - \mu_F\right)^2 + \left(\sigma_F - \rho\sigma_A\right)^2} \rceil + 3$$

Interestingly, the critical value is closely related to the terms of Theil's decomposition (Theorem 2.1). The numerator is the random error component whereas the denominator is the sum of the mean and regression bias components. Consequently, the minimal training sample size is the smaller the lower the random error component in comparison to the systematic components. In other words, a forecast with strong biases requires smaller training samples for the corrected forecast to outperform the original one.

The minimal training sample size under the bivariate normality assumption is illustrated and discussed in Section 3.3 to assess the actual training sample size required, depending on the characteristics of the actuals and forecasts.

It should again be noted that the derived training sample size is only valid for normal distributed actuals and forecasts. However, the actuals and forecasts could follow other distributions such as a t-distribution as an extreme example. In this case, outliers, which are likely to destabilize the regression coefficient estimates, are much more likely. The destabilization results in a larger sampling variance of the coefficient and, as a consequence, a higher minimal training sample size in comparison to the one derived under the normality assumption. Thus, the derived minimal sample sizes should only be seen as an approximate guideline and a substantial number of additional training data points should be used, depending on the distribution of the data.

Another assumption underlying the analysis of the minimal training sample size is that the population parameters of the forecasts and actuals are identical for the training and the evaluation data, i.e., that no structural change occurs. The next analyses focus on the additional effect of these changes.

3.2.2 Structural Change

Structural changes in error patterns and biases are likely to occur in practice, for instance because of learning effects or staff turnover. As biases are assumed to be time-invariant in forecast correction methods, at least some changes can be expected to negatively influence forecast correction. If, for instance, biases completely vanish and future forecasts are completely unbiased, the old biases are falsely removed, which in turn introduces a new systematic bias.

To enable a thorough analysis and discussion of the effect of structural changes, some additional notation has to be introduced.

In contrast to the previous analyses, an analysis of structural changes requires at least two samples: one data sample before and one after the structural break. A model learned from data before and after the structural break is applied to new data from after the structural break. In order to limit the complexity of the analyses, it can be assumed that the structural break occurs between the two samples, i.e., the parameters are estimated from data with old biases and applied to data with changed biases.

The random variables representing the actuals and forecasts are denoted A, F for the training sample and $\tilde{A}, \tilde{F}$ for the evaluation sample. Actuals and forecasts have standard deviations σ_A, σ_F ($\tilde{\sigma}_A, \tilde{\sigma}_F$) and correlation ρ ($\tilde{\rho}$) in the training (evaluation) sample. Based upon these characteristics, the regression slope coefficient can be calculated for both samples as β_1 and $\tilde{\beta}_1$. Given these definitions, the expected MSE in the evaluation sample is defined in Equation 3.5.

$$
\begin{aligned}
MSE =& \mathrm{E}\left[\left(\tilde{A} - \left(\beta_0 + \beta_1 \tilde{F}\right)\right)^2\right] \\
=& \mathrm{E}\left[\left(\left(\tilde{A} - \mu_A\right) - \beta_1\left(\tilde{F} - \mu_F\right)\right)^2\right]
\end{aligned}
\tag{3.5}
$$

If no change occurs between training and evaluation sample, all parameters are equal for both samples and the MSE reduces to the random error, as shown in Theorem 2.2.

In contrast, the MSE is higher than the random error if one of the sample characteristics changes. This can be motivated intuitively since the forecast error is only completely reduced to the random error component if the biases detected in the training sample match the biases in the evaluation sample perfectly (and if the parameter estimation uncertainty is ignored). This is not the case if a change occurs, resulting in an increase of the MSE, which can be analyzed for different types of changes.

First, the mean error (corresponding to the mean bias) can change between training and evaluation sample. Assuming that the mean of the actuals is unchanged, the change of the mean error directly corresponds to a change of the mean of the forecasts. In this case, the MSE increases as presented in Theorem 3.4.

Theorem 3.4 (Influence of Mean Error Change)**.** *Assuming a change of the mean forecast value between the training and the evaluation sample by Δ_μ. Given the coefficient estimate in the evaluation sample, $\tilde{\beta}_1$, the MSE increases in comparison to the in-sample random error by*

$$
\Delta MSE = \left(\tilde{\rho}\frac{\tilde{\sigma}_A}{\tilde{\sigma}_F}\Delta_\mu\right)^2
$$

The change of the mean forecast has quadratic influence on the MSE, which matches the influence of the difference of the mean actual and mean forecast values in Theil's decomposition. Furthermore, the influence is stronger for larger values of $\beta_1 = \tilde{\beta}_1$. This results is also clear as the intercept term in a linear regression uses the estimate of β_1. Larger values of β_1 have consequently stronger influence on the intercept estimate and thus on the error if the mean forecast (and in turn the optimal intercept) changes.

The MSE increase can be used to derive a critical change of the mean forecast value, which results in the corrected forecast performing equal to the original forecast. The critical change is shown in Theorem 3.5.

Theorem 3.5 (Critical Change of Mean Forecast)**.** *Assuming a correction model is applied to an evaluation sample (with $\tilde{\sigma}_A, \tilde{\sigma}_F, \tilde{\rho}, \tilde{\mu}_A, \tilde{\mu}_F$), then the corrected and the orig-*

inal forecast have equal MSE if the mean value of the forecast changed, from the training to the evaluation sample, by

$$\mathring{\Delta}_{\mu} = \pm \frac{\tilde{\sigma}_F}{\tilde{\rho}\tilde{\sigma}_A} \sqrt{\left(\tilde{\mu}_A - \tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A\right)^2}$$

Like for the critical sample size, a close relationship to the terms of Theil's decomposition can be noted for the critical change in Theorem 3.5. The expression under the square root is the bias-related part of the MSE. Consequently, the critical change increases with increasing biases within the evaluation sample. The critical value is, however, also influenced by the inverse of β_1, which introduces an additional dependency on the bias in the forecasts. If the standard deviation of the forecasts is substantially lower than the correlated part of the standard deviations of the actuals (i.e. the variance of the actuals is systematically underestimated by the forecasts), the critical value further decreases.

Second, the correlation between actuals and forecasts can increase or decrease from one sample to the other. This change corresponds to a shift between the random error and the regression bias component. For instance an increase in correlation indicates a reduction of the random error component. The influence of this change of the MSE is analyzed in Theorem 3.6

Theorem 3.6 (Influence of Correlation Change). *Assuming a change of the correlation between forecasts and actuals from the training to the evaluation sample by Δ_ρ. Then the MSE increases in comparison to the in-sample random error by*

$$\Delta MSE = \left(\Delta_\rho\right)^2 \tilde{\sigma}_A^2$$

As can be seen in Theorem 3.6, the change of the correlation has a quadratic influence on the MSE of the corrected forecast. As a change of the correlation corresponds to a shift between regression bias and random error in Theil's decomposition, the strength of the regression bias is over- or underestimated if the correlation changes. Thus, the bias in the evaluation sample cannot be removed completely or the correction is too strong.

The introduced MSE change as a result of the correlation change can be used to derive a critical change of the correlation so that the corrected forecast performs equal to the original biased forecast. This critical change is presented in Theorem 3.7.

Theorem 3.7 (Critical Change of Correlation). *Assuming a correction model is applied to an evaluation sample (with $\tilde{\sigma}_A, \tilde{\sigma}_F, \tilde{\rho}, \tilde{\mu}_A, \tilde{\mu}_F$), then the corrected and the original forecast have equal MSE if the correlation between forecast and actuals changed, from the* *training to the evaluation sample, by*

$$\mathring{\Delta}_\rho = \pm \frac{1}{\tilde{\sigma}_A} \sqrt{(\tilde{\mu}_A - \tilde{\mu}_F)^2 + (\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A)^2}$$

Interestingly, the critical correlation change in Theorem 3.7 resembles the critical mean forecast change in Theorem 3.5 as the same square root (of the sum of the bias components in Theil's decomposition) is included in both formulae. Consequently, the threshold of the correlation change also increases with increasing systematic biases since larger changes are required to negate the advantage of the corrected forecast over the original biased forecast. Additionally, a smaller variance of the actual values decreases the critical value. This effect can be explained by the change of the regression bias. Noting that the variance of the forecasts is fixed and unchanged, a small variance of the actuals corresponds to a strong regression bias where the corrected forecast has a substantial benefit in comparison to the original forecast, which requires larger changes to negate.

Third, the variance of the errors can change. Assuming that the characteristics of the actuals are unchanged, this corresponds to a change of the variance of the forecasts. The forecast variance, like the correlation, influences the regression bias component of the forecast error. For instance if the forecast variance increases, the part of the forecast variance exceeding the variance of the actuals also increases, in turn increasing the regression bias component. The increase of the MSE, which results from a change in forecast variance, is analyzed in Theorem 3.8.

Theorem 3.8 (Influence of Forecast Variance Change)**.** *Assuming a change of the standard deviation of the forecasts by Δ_σ between training and evaluation sample. Given standard deviations of actuals and forecasts $\tilde{\sigma}_A, \tilde{\sigma}_F$ and correlation $\tilde{\rho}$ in the evaluation sample, the MSE increases in comparison to the in-sample random error by*

$$\Delta MSE = \left(\tilde{\rho}\frac{\Delta_\sigma}{\tilde{\sigma}_F - \Delta_\sigma}\right)^2 \tilde{\sigma}_A^2$$

The change of the MSE in Theorem 3.8 is the stronger, the larger the change of the standard deviation of the forecasts Δ_σ in comparison to $\tilde{\sigma}_F$. Clearly, a slight increase of the variance of the forecasts has a smaller impact when the variance is already high. Furthermore, the change depends on the correlation between forecasts and actuals. If the correlation is low, the original forecast has little weight in the corrected forecast and the change of the forecast variance has consequently little impact. In contrast, if (in the training sample) the correlation was high and the forecasts were highly correlated to the actuals, a change of the variance would results in a considerable change of the regression bias.

Using the introduced change of the MSE in case of a change of the forecast variance, a third critical change can be derived, which is shown in Theorem 3.9.

Theorem 3.9 (Critical Change of the Forecast Variance)**.** *Assuming a correction model is applied to an evaluation sample (with $\tilde{\sigma}_A, \tilde{\sigma}_F, \tilde{\rho}, \tilde{\mu}_A, \tilde{\mu}_F$), then the corrected and the original forecast have equal MSE if the variance of the forecasts changed, from the training to the evaluation sample, by*

$$\mathring{\Delta}_\sigma = \frac{\sigma_F \left((\tilde{\mu}_A - \tilde{\mu}_F)^2 + (\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A)^2 \right) \pm \tilde{\rho}\tilde{\sigma}_A \sqrt{(\tilde{\mu}_A - \tilde{\mu}_F)^2 + (\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A)^2}}{(\tilde{\mu}_A - \tilde{\mu}_F)^2 + (\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A)^2 - \tilde{\rho}^2\tilde{\sigma}_A^2}$$

The critical value in Theorem 3.9 can have two solutions which are valid, i.e., correspond to a positive standard deviation of the forecasts in the training sample. Consequently, different increases or decreases of the can be critical. Unfortunately, as a result of the complexity of the critical value formula, a detailed discussion of the influences on the critical value as well as of their strengths is not straightforward. However, Section 3.3 later illustrates the thresholds derived in this section, including the complex variance change threshold.

Clearly, the changes of MSE resulting from the three changes that are studied in this section are not completely independent. The changes presented in Theorems 3.4 to 3.8 are consequently not additive and changes can reinforce or negate each other. Nevertheless, the analyzed influences allow interesting insights into the influence of structural changes on forecast correction methods.

In summary, changes of the sample characteristics and biases clearly affect the performance of forecast correction methods. However, as discussed in two of the three analyses, as long as substantial biases exists in the forecasts, larger changes are in most cases required in order to put the original biased forecast substantially in favor of the corrected forecast. This is a result of the advantage that the corrected forecast can be expected to have in comparison to the original forecast and that requires substantial changes to negate.

Up to this point, estimation uncertainty as well as structural changes, which are likely to influence forecast correction in practice, have been analyzed. However, the analyses for this purpose only focused on the population parameters of the forecasts and actuals while other characteristics of the time series have been excluded. The influence of non-stationarity as an important characteristic of time series is analyzed next.

3.2.3 Non-Stationarity

Systematic components such as seasonal or a linear trend components are likely to occur in time series in practical applications,. These systematic components can for instance be a result of business characteristics such as seasonal cycles or a business growth. In judgmental forecasting, a human expert producing a forecast for a time series with a strong systematic component is likely to notice the systematic component and include it into the forecast. Consequently, the underlying business developments not only influence the actual values of the time series but also the corresponding forecasts.

This relationship is a clear case of a spurious relationship, as introduced by Yule (1926), since the forecast and actual time series are to a certain extent influenced by the same underlying factor. If variables, such as time series, with a spurious relationship are used in a linear regression, the results can be rather unexpected, an effect for which the term spurious regression was coined. See Granger and Newbold (2014) for an overview of the characteristics and effects of spurious regression.

As forecast correction relies on linear regression, spurious relationships as a result of systematic components in time series can be expected to influence the models. This can for instance be seen in the results of the experiments of Goodwin (1997), who included time series with a linear trend into the experiments. For these time series, the results showed that using Theil's method substantially decreased the accuracy of the forecasts when a low noise was present in the actual time series. While Goodwin attributed this observation to how humans forecast time series with a linear trend, the spurious relationship provides an alternative explanation. The effect of a spurious relationship between actuals and forecasts is analyzed in Theorem 3.10.

Theorem 3.10 (Spurious Relationship in Theil's Method)**.** *Let T be a systematic component with variance σ_T^2 and with zero covariance with forecasts and actuals. If T is added to actuals and forecasts, resulting in modified forecasts F' and actuals A', then the regression bias component in Theil's method is influenced towards 0.*

Theorem 3.10 shows that the detectable regression bias vanishes with increasing strength of the systematic component. The strength of the component is for this purpose measured by its variance. Clearly, a stronger linear trend component has a higher variance than a weaker one. Likewise, a strong cyclical seasonal component has higher variance.

As the detectable regression bias vanishes for strong systematic components, the potential reduction of the bias component of the forecast error by using forecast correction also decreases. However, if a forecast correction method is nevertheless applied, the variance component is still increased as a result of the parameter estimation. Thus, in terms of the bias–variance trade-off, a spurious relationship results in a low reduction of the bias component that can be negated by the variance component.

Overall, forecast correction is clearly prone to spurious relationships as a result of non-stationarity of time series. In time series forecasting, where non-stationarity is also an issue, the standard approach is to apply pre-whitening to non-stationary time series, for instance by differencing or detecting and removing the systematic component. Different approaches are for instance discussed by Brockwell and Davis (2013).

Many approaches to pre-whitening time series introduce additional parameters, which have to be estimated from past data. An example is detrending, where a trend, i.e., an on average constant increase per time period, is first estimated and then removed from past actuals and forecasts as well as from future forecasts.

On the one hand, pre-whitening can decrease the bias component of the error of the corrected forecast. If non-stationarity prevents detecting and removing biases in a reasonable way, pre-whitening is an essential basis for the correction approach. However, on the other hand, the estimation for the pre-whitening approach can introduce additional parameters, which results in an increased variance component of the error.

As a consequence of the resulting bias–variance trade-off, pre-whitening should not be applied to all time series. Depending on the time series, the pre-whitening approach (or no pre-whitening approach) should be applied that results in the best trade-off between the bias and the variance components. Complex pre-whitening approaches should only be used if justified by the time series characteristics. The question which approach to use in which case is a main aspect of the empirical evaluation in Chapter 6.

Although the basic effect of non-stationarity is clear from the analysis in this section, the next section gives an example to illustrate the consequences of non-stationarity. Likewise, the results regarding training samples size and structural changes are illustrated to provide additional insights.

3.3 Illustration and Discussion

The analyses in the previous section resulted in a variety of theoretical results regarding the effects of training sample size, structural changes, and non-stationarity of time series. While important influences and characteristics have already been discussed for some of the analytical results in order to derive a basic understanding, this section provides further illustration and discussion of the results. First, the minimal training sample size threshold is illustrated. Subsequently, the critical changes of the different bias-related parameters are analyzed and discussed. Lastly, an example illustrating the effect of non-stationarity of time series on the detectability of biases is given.

3.3.1 Training Sample Size

For a discussion of the minimal sample size required to –in expectation– outperform the original forecast (as introduced in Theorem 3.3 under the assumption of bivariate normality of forecasts and actuals), two assumptions are useful. First, $\sigma_A = 1$, which does not limit generality since the actuals and forecasts can always be scaled in a way to satisfy this assumption. Second, $\mu_A - \mu_F = 0$, which corresponds to a forecast without mean bias. As can easily be seen, the minimal sample size decreases if a mean bias exists. Assuming that no mean bias exists consequently results in a more conservative value of the minimal training sample size.

The resulting minimal sample size is displayed in Figure 3.1 as a function of σ_F and for selected values of the correlation between forecasts and actuals ρ. As can be expected from the formula, the minimal sample size strongly increases if σ_F approaches $\rho\sigma_A = \rho$. This result can be interpreted intuitively since the regression bias vanishes in this case (while there already is no mean bias by assumption). Weak biases make it increasingly harder for the corrected forecast to outperform the original forecast. To outperform the original forecast in case of a very weak bias, a very large sample is required in order to make the estimated parameters very stable. The small reduction of the bias component is only larger than the increase of the variance component if the estimates are very stable.

However, if a substantial bias exists, relatively small training samples suffice for the corrected forecast to outperform the original biased one. Under the assumption of bivariate normality, sample sizes of 10 to 20 are in most cases sufficient, which is a sample size available in many practical applications. Approximate guidelines for cases where a correction is likely to be reasonable could be derived from Figure 3.1. For instance $\left|\frac{\sigma_F}{\sigma_A} - \rho\right| > 0.2$ would cover most cases where a training sample size of 10 to 20 is sufficient.

It should again be noted that the analysis of the minimal training sample size is based upon the assumption of bivariate normality of the actuals and forecasts. While the errors of the forecasts might often satisfy normality, the normality of actuals and forecasts cannot be assumed in general. The required sample sizes

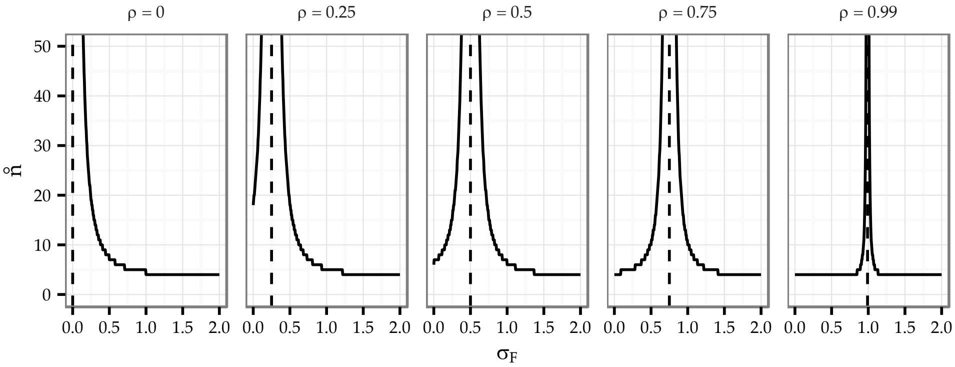


Figure 3.1: Minimal training sample size for the corrected forecast to outperform the original biased forecast without mean bias. The minimal sample size increases for a regression bias approaching zero ($\sigma_F = \rho\sigma_A = \rho$, horizontal dashed line).

required to outperform the original forecast can be substantially larger for other distributions. However, the analysis excluded the mean bias, which would again decrease the required minimal training sample size. Considering both aspects, the guidelines should only be seen as approximate ones and larger sample sizes should be used to ensure sufficient stability.

Overall, the analysis of the analytically derived minimal training sample size in case of bivariate normality of forecasts and actuals implies that relatively small training sample sizes of 10 to 20 suffice for the corrected forecast to outperform the original one. This is especially the case if the original forecast is strongly biased, which results in a strong advantage of the corrected forecast and makes the outcome less prone to estimation errors. Thus, considering the required training sample size, forecast correction can be expected to be successful in many settings.

3.3.2 Structural Change

Although relatively small training sample sizes are often sufficient, a low robustness against structural changes can also limit applicability. To discuss this aspect, the derived critical values regarding structural changes can be analyzed.

The first analyzed threshold is the change of the mean error as a result of a change of the mean value of the forecasts, which was introduced in Theorem 3.5 depending on the characteristics of the evaluation sample. The critical value answers the question how strongly the mean error must have changed from the training to the evaluation sample if the corrected forecast and the biased forecast perform equal in the evaluation sample. The critical change of the mean value of the forecasts $\mathring{\Delta}_\mu$ is displayed in Figure 3.2 as a function of the mean bias in the evaluation sample and for different values of $\sigma_F = \tilde{\sigma}_F, \rho = \tilde{\rho}$. It should be noted that only the positive critical value is shown. Additional negative critical values exist that are however symmetrical to the positive ones, as can be seen in Theorem 3.5.

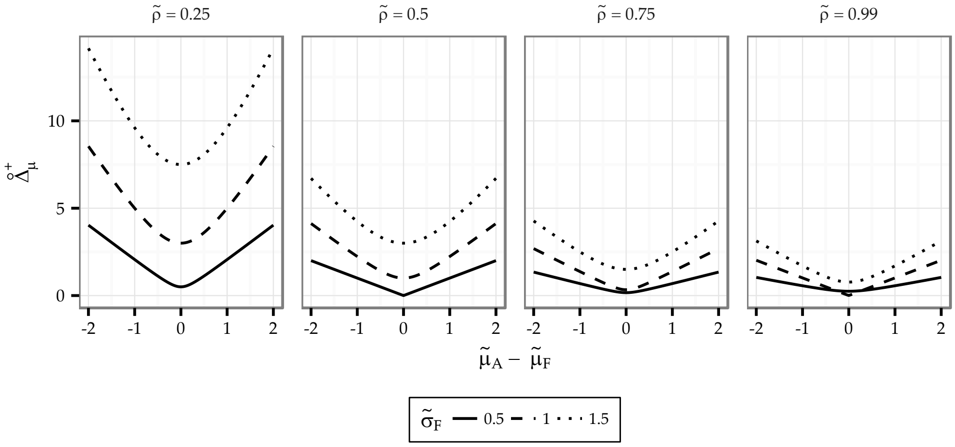


Figure 3.2: Positive critical changes of the mean forecast value for different $\rho = \tilde{\rho}$ and $\sigma_F = \tilde{\sigma}_F$. The critical change increases with increasing mean bias of the forecast and with increasing regression bias, i.e., higher σ_F or lower ρ. Negative changes are symmetric to the positive ones.

The critical change of the forecast mean clearly increases with increasing mean bias in the evaluation sample. If a bias exists, the corrected forecast has a substantial advantage over the original forecast, which requires a large change of the bias to negate the effect. This effect is even stronger if the correlation between forecasts and actuals is very low or if $\sigma_F = \tilde{\sigma}_F$ substantially exceeds $\sigma_A = 1$. In these cases, the corrected forecast can remove a substantial regression bias from the forecast, which results in lower errors and in turn requires larger changes of the mean bias for the MSE of the two forecasts to be equal.

Figure 3.3 displays the critical change of the correlation between forecasts and actuals, introduced in Theorem 3.7, as a function of the correlation $\tilde{\rho}$ and for different values of the mean bias and forecast standard deviation $\sigma_F = \tilde{\sigma}_F$. The critical change again indicates how strongly a change must have been if the original and the corrected forecast perform equal in the evaluation sample.

The figure illustrates a strong relationship between the critical change and the mean bias. The critical value increases with the strength of the mean bias. If the mean bias is strong enough ($\tilde{\mu}_A - \tilde{\mu}_F = 2$ in the figure) critical changes do not exist, i.e., would require correlations with absolute value larger than one. The critical change is, independently of the mean bias, lowest for $\tilde{\sigma}_F = \tilde{\rho}, \tilde{\sigma}_A = \tilde{\rho}$ as

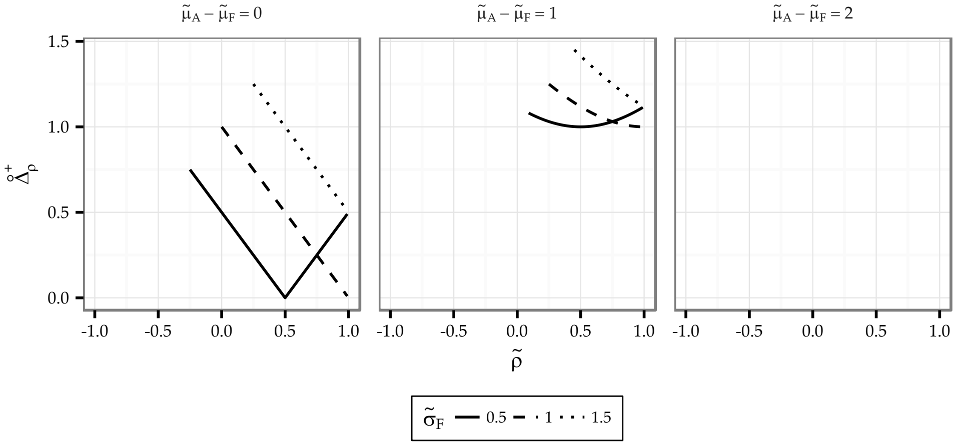


Figure 3.3: Positive changes of the correlation between actuals and forecasts for different values of $\mu_A - \mu_F = \tilde{\mu}_A - \tilde{\mu}_F$ and $\sigma_F = \tilde{\sigma}_F$. The critical changes increase with the strength of the mean bias. If the mean bias is strong enough, critical changes do not exist, i.e., would require correlations with absolute value greater than one. The critical change is lowest for $\tilde{\sigma}_F = \tilde{\rho}, \tilde{\sigma}_A = \tilde{\rho}$ where the regression bias vanishes.

the regression bias vanishes in this case. If both biases are low, the critical value decreases towards zero. This confirms the previous discussion of the analytical thresholds, which indicated that stronger biases increase robustness against changes.

Lastly, Figure 3.4 presents the critical changes of the standard deviation of the forecasts introduced in Theorem 3.9 as a function of the standard deviation in the evaluation sample $\tilde{\sigma}_F$ and for different values of the mean bias and correlation between forecasts and actuals $\rho = \tilde{\rho}$. Like in the previous analyses, the critical change indicates how strongly the standard deviation must have changed if the corrected and original forecast have equal MSE.

As can be expected from the complexity of the threshold formula in Theorem 3.9, the critical changes are rather complex, especially if the mean bias is zero. Two aspects can immediately be noted in the case without mean bias. First, increases as well as decreases can be critical in some cases. The standard deviation of the forecast can influence the regression bias in both directions, which can both be critical, however at a different level. Second, the correction has the lowest robustness if the regression bias is also close to zero, i.e., for $\tilde{\sigma}_F = \tilde{\rho}, \tilde{\sigma}_A = \tilde{\rho}$. If a mean bias exist, the critical changes are much higher and relatively systematic, except for $\tilde{\mu}_A - \tilde{\mu}_F = 1$ and $\tilde{\rho} = 0.99$. Thus, the critical values regarding changes of the variance of the forecast conforms to the results of the previous analyses: stronger biases increase robustness against changes.

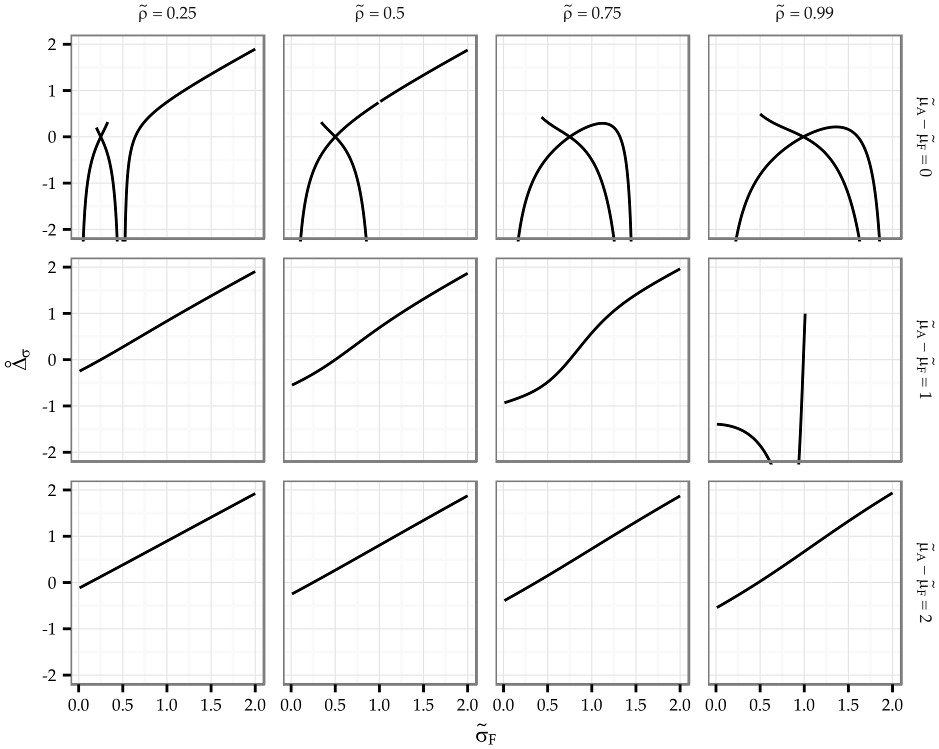


Figure 3.4: Positive and negative changes of the standard deviation of the forecasts for different values of $\mu_A - \mu_F = \tilde{\mu}_A - \tilde{\mu}_F$ and $\rho = \tilde{\rho}$. Critical changes are smallest if no mean bias exists (upper row), especially if the regression bias is also small, i.e., $\tilde{\sigma}_F = \tilde{\rho}, \tilde{\sigma}_A = \tilde{\rho}$. A mean bias quickly leads to considerable robustness.

Overall, the robustness against structural breaks can in some cases be very low with respect to all three influences under study. Since this was always found to be the case if mean and regression bias are very low, i.e., if the forecast is close to unbiasedness, biases in the forecasts result in robustness against changes. If biases exist, the bias es must be strongly reduced or disappear for the original forecast to outperform the corrected one. In other words, if only a small bias existed in the original forecast (and the potential advantage is consequently very small) and the bias additionally change in a disadvantageous way, the corrected forecast can have worse accuracy than the original forecast. Thus, if large biases exist, a correction can be applied with higher confidence regarding error improvements than in cases with low biases.

3.3.3 Non-Stationarity

Whereas biases increase robustness against changes, the analyses indicated very different influence of non-stationarity of time series on forecast correction. Strong non-stationarity of time series results in vanishing biases, which not only decreases robustness against changes but also decreases the potential bias reduction effect of forecast correction.

To illustrate the effect of non-stationarity of time series, an illustrative simple example is shown in Figure 3.5 for a simple example. In the left part of the figure, an actual time series and a corresponding forecast time series is displayed in the upper plot. The actuals and the forecasts clearly have a shared trend component that causes the strong increase of actuals and forecasts. The lower left plot displays the corresponding forecast-actual plot and the estimated regression coefficients. The estimates indicate that the forecasts are nearly unbiased, with $\hat{\beta}_1 = 0.99$ close to 1 and $\hat{\beta}_0$ close to 0 (considering the range of the actuals and forecasts). The forecast-actual plot alone indicates that the forecasts are reasonably unbiased.

Since actuals and forecasts exhibit a shared trend component, the upper right part of the figures shows the time series after removal of the trend component. Actuals as well as forecasts fluctuate randomly and the previously clear relationship between forecast and actuals vanishes. This is even clearer for the forecast-actual plot of the detrended forecasts and actuals in the lower right part of the figure. The forecasts in the example contain no information regarding the actual values (since $\hat{\beta}_1 = 0$) beyond the shared trend component.

The example clearly illustrates the influence of a shared systematic component in forecasts and actuals. Although the example is an artificial and extreme one, the effect can be expected to generalize well even though forecasts are in practice (and in contrast to the example) likely to contain information beyond the shared component.

Thus, a systematic component in the time series, which results in non-stationarity, is a clear issue for forecast correction. If for instance trended time series are corrected, the detectable bias –and consequently the error reduction– is very weak whereas the error is increased by the required parameter estimation.

This result motivates that not only robustness against structural changes, but also an adequate treatment of non-stationarity of time series might be essential for a successful application of forecast correction. Both aspects are considered in an extended forecast correction model in the next section.

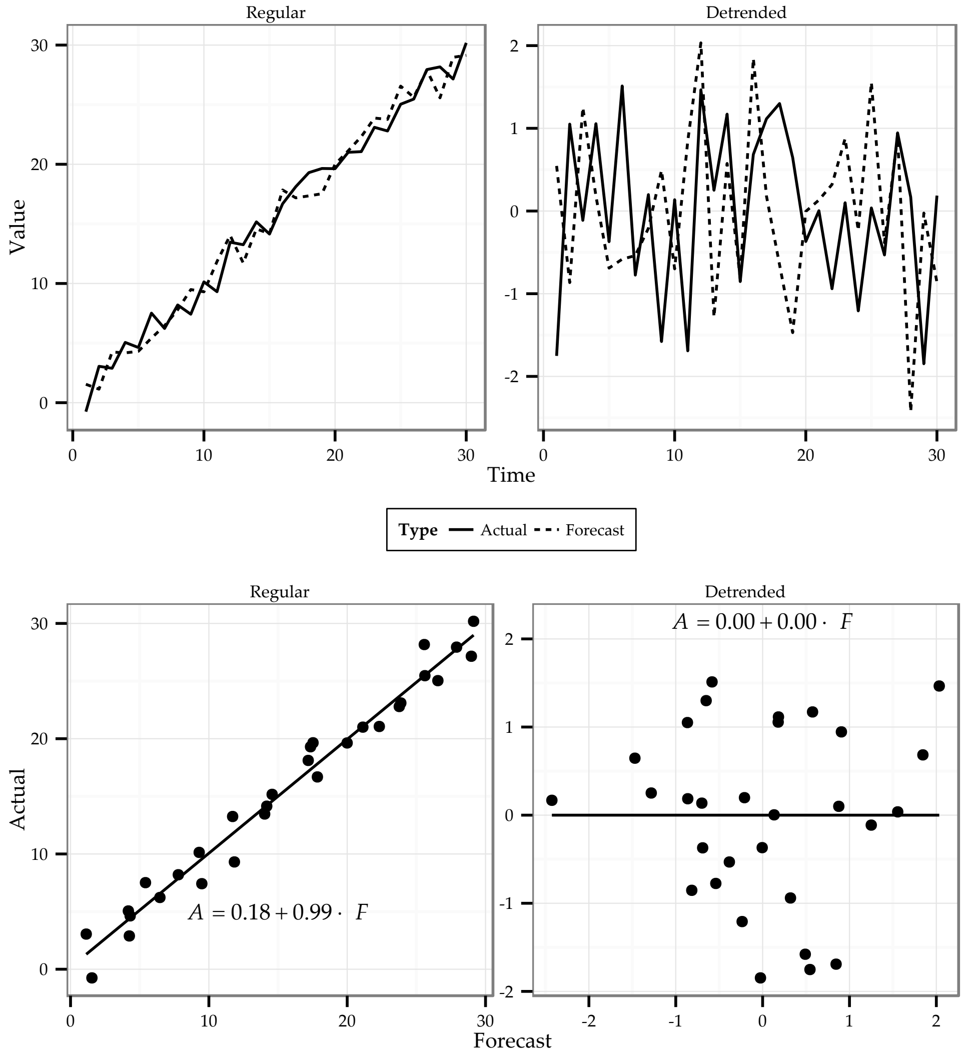


Figure 3.5: Example of the effect of a spurious relationship between actuals and forecasts. While the original time series show a clear relationship between forecasts and actuals and the forecast-actual plot indicates unbiasedness (left part of the figure), a detrending of the time series reveals that the forecasts contain no information on the actuals beyond the shared trend component (right part of the figure).

3.4 Correction Considering Structural Change and Non-Stationarity

The analyses and discussions in the previous sections allowed various insights into the robustness of Theil's method. The results indicated that a corrected forecast can be expected to outperform the original forecast in many cases even when only a small training sample is available. In contrast, structural breaks, resulting in changed biases, can often lead to the original biased forecast performing better than the corrected forecast. Furthermore, non-stationarity of time series decreases the biases that are identifiable with forecast correction models.

In order to transfer these analytical findings to real-world situations with changes and disruptions regarding biases as well as with time series with various characteristics, forecast correction models have to be adapted. The aim is to use the analytical results to change the estimation procedure in a way that the estimated parameters fit future, unknown data better.

First, regarding structural breaks, the existing weighted approach introduced by Goodwin (1997) is one promising approach. However, as Figure 3.6 illustrates, the exponential weighting approach must not necessarily be optimal in case of a structural break. The figure displays an example of exponential and equal weights for a time series with length 36 and a breakpoint at time 18. The discount factor for exponential weights is set to a moderate value of $\gamma = 1.1$.

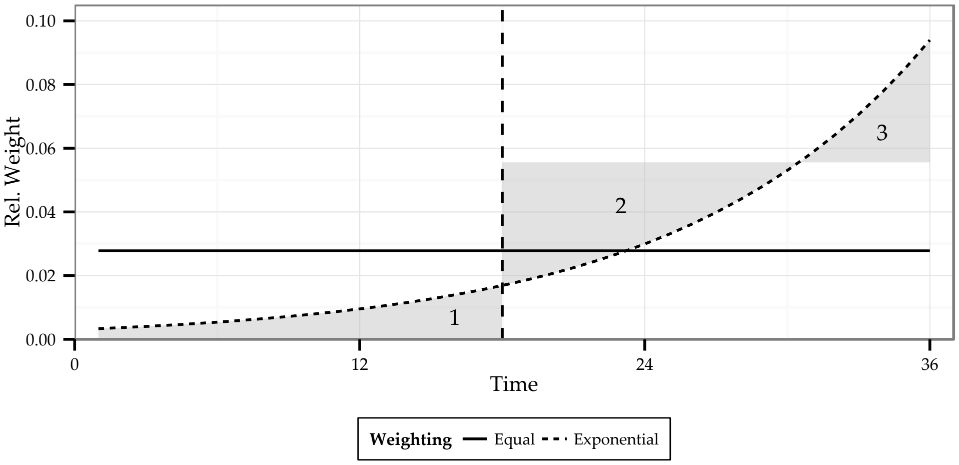


Figure 3.6: The weight distributions for equal and exponential weighting of observations differ strongly. Assuming a breakpoint (dashed vertical line), equal weights assign high weight to observations before the breakpoint. Exponential weights aim at minimizing the weight of outdated observations, but have a very unequal distribution of the weight of the observations after the breakpoint.

Theil's method assigns equal weights to all observations and can clearly be expected not to perform well since the observations before the breakpoint have the same weight as the observations after the breakpoint.

In contrast, the weight function smoothly increases for exponential weights. This weighting scheme assigns far less weight to the observations before the breakpoint. Although the weight of old observations is reduced, weight is still wrongly assigned to these observations (gray area 1). The weight after the breakpoint is furthermore not distributed equally. While weights below equal weights are assigned to some observations after the breakpoint (gray area 2), the weights of the most recent observations by far exceed equal weights (gray area 3). Using exponential weights consequently involves a trade-off between reducing the weights before a breakpoint (corresponding to a reduction of the bias component) and the equality of the weights after a breakpoint (influencing the variance component).

Addressing this trade-off clearly requires information about a potential breakpoint. Since Theil's method is a linear regression approach, techniques from regression theory aiming at detecting and dating structural breaks can also be used in Theil's method. These approaches for detecting breakpoints in linear regression have received much attention in statistics; see for instance Zeileis et al. (2003) for an overview of different approaches.

In principle, detecting the points of time of structural changes (henceforth called breakpoints) corresponds to finding significantly differing partitions of the data separated by the breakpoints. Partitions of the data with a set of m breakpoints $\tau_1, \ldots, \tau_m$ can be identified using Equation 3.6 where RSS is the residual sum of squares resulting from applying separate regressions to the $m + 1$ partitions separated by the breakpoints.

$$\underset{\tau_1,\ldots,\tau_m}{\arg\min}\, RSS\,(\tau_1, \ldots, \tau_m) \tag{3.6}$$

Clearly, RSS must decrease for increasing numbers of breakpoints. Each breakpoint adds parameters to the model and consequently results in a closer fit to the data. In an extreme case, the data could be segmented in a way so that at most two data points are contained in a segment. In this case, the model would consist of many linear regression models for at most two data points, which must always fit the data perfectly. Such extreme cases are obviously not beneficial, especially if the model is later applied to unknown data. As a consequence, the number of breakpoints in a model must be penalized when comparing the RSS results for different number of breakpoints.

The Bayesian information criterion (BIC) proposed by Schwarz (1978) introduces a penalization of the number of parameters of a model and allows a comparison of models with different numbers of parameters. The number of breakpoints (and the corresponding partitions of the data) that minimize the BIC can then be selected.

After a set of potential breakpoints has been identified, the key question is how to incorporate the detected breakpoints into the estimation of a forecast correction model. Since the training data is segmented along the time axis, no reasonable model with multiple regression lines for different segments, which might be appropriate in other applications, can be used in forecast correction models. Clearly, the most important segment is the last segment, which should always be used for parameter estimation since the data is the most recent and consequently most likely to include biases that are identical to those in future forecasts.

In contrast to the last segment, the treatment of other segments is not straightforward. On the one hand, older segments can be ignored if the last segment is large enough and results in a stable model. On the other hand, a small last segment can result in an unstable model and in turn an increased variance component of the error of the corrected forecast. In this case, including older segments might decrease the variance component much more than the increase of the bias component by the old training data with outdated biases. Clearly, the old training data should still not have the same weight as more recent observations and a reduction of the weights must be considered.

A model meeting these requirements is introduced in Equation 3.7 based on the model in Equation 3.4 and using the last detected breakpoint τ_m.

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\arg\min} \sum_{t=1}^{\tau_m-1} \alpha\gamma^t p\left(A_t - \left(\hat{\beta}_0 + \hat{\beta}_1 F_t\right)\right) + \sum_{t=\tau_m}^{T} \gamma^t p\left(A_t - \left(\hat{\beta}_0 + \hat{\beta}_1 F_t\right)\right) \tag{3.7}$$

The minimization in the extended estimation model involves two sums, one with the observations before the last breakpoint and one with those after. In both cases, the errors are again penalized with a function p such as OLS. An exponential discount with parameter $\gamma \geq 1$ is again included in both sums to allow a slow decrease of the influence of older observations. The first sum with the penalized deviations for the observations before the last breakpoint additionally includes a parameter $\alpha \in [0, 1]$, which can be used to further reduce the weight of the observations before the breakpoint. If no breakpoint is detected, only the second term of the model is used with $\tau_m = 1$.

While the motivation for and effect of the exponential weighting parameter γ is clear from the discussion in Section 3.1, the new parameter α requires additional explanation and motivation. If the last breakpoint is at an early point of time, low values for the parameter α result in obsolete observations before the breakpoint having negligible influence, consequently ensuring a low bias component of the error while stable estimates of the parameters and a low variance component of the error are likely because of the high number of remaining observations. In contrast, if the breakpoint occurred at a late point of time, higher values for the parameter α stabilize the estimates by including more observations.

A special case of the introduced model is the simple strategy of ignoring the observations prior to the last detected breakpoint. This strategy can be implemented by simply setting $\alpha = 0$, which results in zero influence of the data points before the breakpoint. Another simple strategy is ignoring the breakpoint and only relying on the exponential weighting for the treatment of the breakpoint. This strategy corresponds to $\alpha = 1$.

However, in contrast to these simple strategies, where the value of α is fixed, a more differentiated treatment of breakpoints requires learning the parameter from past observations. A promising approach is to learn the parameter using the same procedure used for choosing a value for γ: the pseudo out-of-sample evaluation on the last observations of the training data introduced in Section 3.1. Instead of running the same procedure separately for α and γ, the procedure can be executed for combinations of the two parameters in order to identify the optimal combination of values of the two parameters.

Overall, if breakpoints are detected correctly (or at least reasonably accurate), the extended approach can be expected to perform well by balancing model stability and an increased bias component as a result of outdated observations. This promises more robust parameter estimates and an increased out-of-sample performance.

As the estimation model in Equation 3.7 allows a differentiated treatment of structural breaks, which were identified to be important in many cases in the theoretical analyses, an additional extension addressing non-stationarity of time series can be considered. For this purpose, some approaches to ensuring stationarity of time series are discussed first. For this discussion, the forecast horizon h is, in contrast to the previous analyses and discussions, explicitly used since the approaches mostly rely on past data, which is only available up to A_{t-h} when producing a forecast F_t.

A simple approach is to differentiate the available observations, i.e., to use the realized and predicted changes from the last available actual value. This corresponds to the function introduced in Equation 3.8, where X_t is an (available) actual or forecast value and A are all available actual values in the training data. Differentiation in many cases results in stationarity since for instance values fluctuating around a constant trend results in a stationary fluctuation around a constant value, the slope of the trend.

$$\xi_D \left(X_t, A \right) = X_t - A_{t-h} \tag{3.8}$$

Another simple approach, which is common especially in the domain of financial forecasting, is using logarithmic returns, which corresponds to Equation 3.9. The approach also removes a trend from the data. In comparison to differentiation, logarithmic returns also take into account that values are likely to have larger fluctuations for higher levels by using relative changes in the formula. It should be noted that logarithmic returns are not defined if $A_{t-h} = 0$ and should consequently not be applied if zero values are likely to occur in the time series.

$$\xi_L\left(X_t, A\right) = \log \frac{X_t}{A_{t-h}} \tag{3.9}$$

Another approach that can be used to remove a trend from a time series is to explicitly detect and remove the trend. A linear trend in the actual values can be detected by regressing the actual values on the time, i.e., $A = \beta_0^{trend} + \beta_1^{trend}\vec{t} + \epsilon$ where $\vec{t} = (1, \dots, t)$. The estimate $\hat{\beta}_1^{trend}$ is then the average linear trend in the actual time series whereas $\hat{\beta}_0^{trend}$ is a constant offset. Defining $trend_L(A, t) = \hat{\beta}_0^{trend} + \hat{\beta}_1^{trend} t$ as the value of the linear trend component a time t allows defining the trend removal as shown in Equation 3.10.

$$\xi_T\left(X_t, A\right) = X_t - trend_L(A, t) \tag{3.10}$$

The previous approach assumes a constant linear trend in the time series. However, a trend can change in many practical settings. Furthermore, time series can also exhibit a seasonal component where deviations from a trend (or a constant value) occur in a yearly pattern. Both aspects can be addressed in more complex approaches such as the decomposition of time series into seasonality and trend components by Loess (STL), which was introduced by Cleveland et al. (1990). STL results in a decomposition into a trend component $trend_{NL}(A, t)$, which can deviate from linearity, and a seasonal component $seas(A, t)$. These components can be used, as defined in Equation 3.11, similarly to the linear trend in the previous approach.

$$\xi_{TS}\left(X_t, A\right) = X_t - trend_{NL}(A, t) - seas(A, t) \tag{3.11}$$

The different approaches addressing non-stationarity of time series have different properties in terms of the bias–variance trade-off. The basic approach, not addressing non-stationarity, can result in a forecast correction model strongly deviating from the optimal one and consequently introduce a substantial bias component. In contrast, using pre-whitening can reduce the bias component but can also, depending on the approach, introduce additional parameters that in turn increase the variance component. This is particularly true for complex pre-whitening approaches that rely on detecting a systematic trend or even more complex systematic components within forecasts and actuals.

After the approaches to ensuring stationarity of time series have been introduced, the model in Equation 3.7 can be further extended as presented in Equation 3.12, where the actual as well as the forecasts in the training data are transformed, i.e., pre-whitened, for the estimation process.

$$\begin{aligned}\arg\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{t=1}^{\tau_m-1} \alpha\gamma^t p\left(\zeta(A_t,A)-\left(\hat{\beta}_0+\hat{\beta}_1\zeta(F_t,A)\right)\right) \\ +\sum_{t=\tau_m}^{T} \gamma^t p\left(\zeta(A_t,A)-\left(\hat{\beta}_0+\hat{\beta}_1\zeta(F_t,A)\right)\right)\end{aligned} \tag{3.12}$$

After a model has been estimated, the resulting model can be used for correcting a forecast of a future value. However, in contrast to established approaches, a forecast cannot be corrected directly and the corrected forecast cannot be used as calculated by the model. First, the forecast to which the correction model is applied must be transformed in the same way as the actuals and forecasts in the training data. For differentiation and log returns, this step is straightforward. Second, the transformation has to be reversed in order to derive a final corrected forecast. The reversal of the transformation is simple for differentiation and log returns; the corresponding functions are presented in Equations 3.13 and 3.14.

$$\zeta_D^{-1}\left(F_{T+h},A\right)=A_T+F_{T+h} \tag{3.13}$$

$$\zeta_L^{-1}\left(F_{T+h},A\right)=A_T e^{F_{T+h}} \tag{3.14}$$

However, if a trend (or trend and seasonality) is explicitly detected and removed from the time series, the values for the trend and the seasonality component do not exist for future values. As a consequence, the future values of the components have to be predicted. If a linear regression is used for a linear trend detection, the time can simply plugged into the model to derive an extrapolation of the trend component. The resulting linear trend component, denoted $trend_L(A,T+h)$, can then be used to derive a forecast value as presented in Equation 3.15.

$$\zeta_T^{-1}\left(F_{T+h},A\right)=trend_L(A,T+h)+F_{T+h} \tag{3.15}$$

If more complex approaches such as STL are used, more advanced extrapolation approaches have to be used since there is not always a linear relationship between the time and the trend or seasonality. A promising approach is to use time series forecasting methods to extrapolate the values of the components. For instance exponential smoothing can be used for the extrapolation of the trend component while the value of the seasonal component is assumed to be identical to the previous seasonal cycle. The resulting extrapolated trend and seasonality components $trend_{NL}(A, T+h)$ and $seas(A, T+h)$ can then be used for deriving a forecast value as displayed in Equation 3.16.

$$\xi_{TS}^{-1}\left(F_{T+h}, A\right) = trend_{NL}(A, T+h) + seas(A, T+h) + F_{T+h} \tag{3.16}$$

Using the introduced methods and the estimated model parameters $\hat{\beta}_0, \hat{\beta}_1$, the final forecast is defined as presented in Equation 3.17.

$$F_C = \xi^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 \xi(F_t)\right) \tag{3.17}$$

It should be noted that this approach, including the transformation, must also be used in the pseudo out-of-sample evaluation used for determining the values of the parameters α and γ. Consequently, a transformation of the data points prior to each evaluation data point in this procedure is required in a first step. Then, models with different parameter values are applied. In a last step, the correction models are applied and the transformation of the forecast is reversed. After all observations in the pseudo out-of-sample evaluation are treated, errors are calculated and aggregated and a parameterization is chosen. Breakpoints also have to be detected on the transformed actual and forecast values.

In summary, after the theoretical properties of forecast correction approaches have been studied in the previous sections, an extended forecast correction model has been proposed in this section, which is based upon the findings of the theoretical analyses. As structural changes as well as non-stationarity were identified to be relevant issues, the extended model allows a flexible treatment of both aspects. The extended model and its parameterization are evaluated in Chapter 6. In the evaluation, a special focus is placed on which approach to ensuring stationarity should be used in which cases and how robustness against structural changes can be achieved best in practical applications.

3.5 Conclusions and Limitations

Forecast correction is an established approach to improving the accuracy of judgmental forecasts by identifying biases in past forecasts and removing them from new ones. In this chapter, the theoretical properties of forecast correction methods and the influence of three different issues on the accuracy of a corrected forecast were analyzed. First, an analysis of the minimal training sample size required for the corrected forecast to outperform the original biased one revealed that small training samples are unlikely to be an issue for forecast correction. Second, structural changes were shown to have a substantial influence, which can result in the original forecast outperforming the corrected one, especially if the removable biases are already relatively weak. Third, non-stationarity of time series was shown to result in decreased detectability of existing biases, which can in turn result in the correction decreasing forecast accuracy.

As a consequence, an extended model was proposed in order to transfer the analytical results into applications in practice. In contrast to existing forecast correction models, the proposed model explicitly addresses structural changes and considers transformations of actuals and forecasts that ensure stationarity. Structural changes are considered by including breaks detected by established statistical methods into the estimation procedure. Additionally, the exponential weighting of past observations proposed in the literature can be used to address continuously changing biases in the forecasts.

The extensions aim at finding a reasonable trade-off between decreases of the bias component (treating non-stationarity and structural breaks appropriately) and the variance component (estimating additional parameters for data transformation and structural break detection).

While the theoretical analyses revealed that structural changes and non-stationarity are relevant issues, the analyses cannot provide insights into how to address the issues best. The proposed extended model allows various different parameterizations, some of which might be more appropriate than others. In addition, a dependency on the characteristics of the time series can be expected, especially for the data transformation approach, where some approaches might introduce too much uncertainty (such as detecting and removing trend and seasonality from stationary time series).

Overall, additional analyses using real-world forecast data are required to derive guidelines when to use which model and parameterization. A case study evaluating the different methods and aiming at identifying advantageous forecast correction models in a differentiated is presented in Chapters 5 and 6.

The analyses and discussions on forecast correction in this chapter are subject to several limitations. The limitations can be differentiated into general ones, which relate to applying Theil's method for forecast correction, and others that are specific to the analyses in this chapter.

Regarding limitations of Theil's method as the standard approach to correcting judgmental forecasts, an important limitation is that the method relies on the availability of forecast and realization data. Especially if time series are short or past forecasts were not recorded, forecast correction methods cannot be applied. Theil's method is furthermore restricted to detecting and removing linear biases. However, biases might be non-linear in practice and have more complex patterns, which are not considered in the framework of Theil's decomposition.

As Theil's method uses a linear regression, some of the assumptions of linear regression might be violated, which can make the estimates more unreliable. For instance heteroscedasticity might be present in practical applications as forecast errors are likely to have higher variance for higher actuals values. This issue is however not addressed by Theil's method or any of the extensions, including the one introduced in this chapter. Furthermore, the forecast errors, which influence the residuals in the linear regression, might be correlated, for instance as a result of seasonal time series and error patterns.

Regarding limitations of the analyses in this chapter, a first aspect is that the derivation of the minimal training sample size and the conclusions drawn from the analysis depend on the assumption of bivariate normality of the forecasts and actuals. Forecast and actuals must however not always follow a normal distribution in practice. Especially if the distribution of forecasts or actuals has heavy tails, the minimal training sample size required to reasonably learn a forecast correction model might be considerably larger.

The derived critical changes of the biases and forecast characteristics only considered independent changes of individual characteristic. However, structural changes are in practice likely to affect several characteristics. For instance a change of the expert responsible for the forecast not only changes error variance but also the correlation between forecasts and actuals. Simultaneous changes of multiple characteristics can interact, which can reinforce or dampen the influence of the change and result in lower or higher critical changes. The analyses additionally only considered bias changes whereas characteristics of the time series for which forecasts are produced are also likely to be of influence. Actual values might strongly decrease or increase, for instance as a result of business-related changes. Furthermore, the predictability of the time series, and consequently the error variance, can change over time.

However, notwithstanding these limitations, the analyses in this chapter provided various insights into the robustness of forecast correction, which has not been formally analyzed in the literature before. The empirical evaluation presented later in this work analyzes how well the model extensions derived from the theoretical results perform in practice, where the assumptions and limitations are not always satisfied.

Chapter 4

Advances in Forecast Combination

THE combination of forecasts is an established method for improving predictive accuracy. In a combination, errors of individual forecasts cancel each other out and the influence of high errors of one of the forecasts is reduced. Under certain conditions, the combined forecasts can even have lower errors than the most accurate forecast. In economics, the combination of forecasts has been subject to research since the pioneering work of Reid (1968) and Bates and Granger (1969). Numerous studies have shown that the combination of forecasts typically results in increased accuracy in comparison to individual forecasts (Makridakis et al., 1982; Clemen, 1989; Makridakis and Hibon, 2000).

Similar results exist in social psychology, where the wisdom of the crowd phenomenon refers to the fact that an aggregate of judgments often performs better than the best-performing individual, which has been demonstrated in numerous studies (for the benefits of judgment aggregation, see for instance Hill (1982), Hastie (1986), Wallsten et al. (1997), Gigone and Hastie (1997), Hastie and Kameda (2005), or Kerr and Tindale (2011)).

The key questions that are subject to ongoing research are (i) how many and which forecasts to include and (ii) which combination mechanism, i.e., weighting scheme, to use for a combination. To address these questions, the literature on forecast combination as well as on crowd wisdom provides qualitative guidelines and recommendations mainly derived from empirical results.

A clear relationship to the bias–variance trade-off exists for both questions. Most forecast combination methods involve learning the weights of the forecasts from past forecasts and corresponding errors. With each additional forecast included in a combination, additional parameters have to be estimated and the variance component increases. The bias component on the other hand decreases because of the better fit to the data, but the decrease might be negligible in comparison to the increase of the variance component. Likewise, different combination methods are likely to have differing properties in terms of the bias–variance trade-off as the sensitivity to the training data differs. Furthermore, combination methods are likely to be prone to structural breaks in the error patterns to a different degree, depending on the sensitivity.

While the relationship to the bias–variance trade-off is clear, the exact nature of the trade-off and how it is best addressed is unknown. In order to address this issue, Section 4.1 first reviews the theory and literature on forecast combination. Section 4.2 then analyzes the theoretical properties of a class of forecast combination methods. The identified properties are used in Section 4.3 to determine the robustness of a decision on a forecast combination method against small training sample sizes as well as structural changes regarding error covariances.

In Section 4.4, two new combination methods are introduced using the previous analytical results. The optimal shrinkage level optimally solves the bias–variance trade-off involved in forecast combination whereas the robust shrinkage factor ensures robustness against structural breaks to a certain extent.

The results of the theoretical analyses of the robustness and the introduced shrinkage levels are illustrated and discussed in Section 4.5. Finally, Section 4.6 concludes and discusses the implications and limitations of the results.

4.1 Theory and Issues of Forecast Combination

The two combination methods previously introduced in Section 2.2, the simple average (SA) and optimal weights (OW), are completely different and extreme approaches in terms of the bias–variance trade-off. OW aims at minimizing the bias component and consequently results in lower in-sample error variance than any other weighting approach while the variance component of the error of the combined forecast is ignored. In contrast, SA ignores the training data and thus does not aim at minimizing the bias component. However, the weights are fixed, which eliminates the variance component.

Since SA and OW as extreme approaches in terms of the bias–variance trade-off do not necessarily result in minimal out-of-sample error variance, alternative weight estimation approaches have been proposed and evaluated. These for instance include variants of optimal weights constrained to the interval $[0, 1]$, shrinkage towards the average, and Bayesian outperformance probabilities. Clemen (1989), Diebold and Lopez (1996), Armstrong (2001), and Timmermann (2006) provided thorough literature reviews and guidelines regarding the various approaches to forecast combination. A surprising observation of these reviews is that other approaches typically do not outperform SA in out-of-sample evaluations and that weights should only be learned in case of strong evidence against equal weights. See for instance Aksu and Gunter (1992), Stock and Watson (1999), Stock and Watson (2004), and Genre et al. (2013) for studies with this result. Stock and Watson (2004) coined the term "forecast combination puzzle" for this result.

In judgment aggregation, simple rules for aggregating judgments (such as median or mean) are also regularly found to perform at least as good as complex strategies. Mannes et al. (2014) proposed to use averaging after the top judges have been selected. Genre et al. (2013) observed that averaging expert forecasts of unemployment rate and GDP growth typically outperforms the best individual forecast as well as more complex combination schemes. Although the average was outperformed in some cases, the authors cautioned against any assumption that the improvements in these cases would persist in the future.

For the case with two forecasts, differentiated guidelines on which forecast combination method to choose have been proposed based on theoretical and empirical results. Schmittlein et al. (1990) recommended SA combination when errors have similar variances and are only weakly correlated, or for small training samples. More concretely, for sample sizes of ten or below, SA was recommended when error standard deviations differ by at most 20 % and the error correlation is between -0.6 and 0.6. For larger sample sizes of 25 and more and with error correlation between -0.4 and 0.4, error standard deviations must not differ by more than 10 %. However, Schmittlein et al. (1990) did not derive guidelines for more than two forecasts but instead proposed a heuristic using the Akaike Information Criterion (AIC) to compare the different combination methods.

Similarly, de Menezes et al. (2000) recommended SA only for approximately equal error variances. For unequal error variances, medium or large training samples, and error correlations over 0.5, the authors proposed optimal weights constrained to the interval $[0, 1]$. This form of regularization is applied to prevent extreme weights, which can potentially result in increased combined errors. For error correlations below 0.5, outperformance probabilities were recommended for small sample sizes, OW with independence assumption (i.e. using an assumed correlation of zero instead of the estimated correlation) was recommended for medium sample sizes, and OW was recommended for large samples. Thresholds for the similarity of error variances or sample sizes separating small, medium, and large samples were however not quantified.

Overall, the guidelines recommend SA in various cases, except for when strong evidence for another weighting scheme exists. This recommendation and the forecast combination puzzle can be considered surprising as SA does not use available training data, which could in principle be used for learning weights.

From a statistical perspective, the combination weighting schemes and consequently the forecast combination puzzle can be analyzed on the basis of the bias–variance trade-off. SA as a simple approach does not learn from the training data and can consequently have a considerable bias component while a variance component does not exist. In contrast, complex mechanisms have many parameters and are thus highly sensitive to the training data, which results in a low bias component and an increased variance component. The forecast combination puzzle indicates that either the reduction of the bias component in comparison to SA is negated by the increase of the variance component or that the reduction of the bias component is unexpectedly low. Two issues can be identified from the literature as potential causes of these effects.

On the one hand, structural changes of time series characteristics or in the forecasting process are likely to entail changes in the error covariance matrix (which is the basis of weight estimation) over time. In such cases, parameters are differing within the training sample and also between training sample and evaluation sample. This difference in turn results in an unexpectedly low decrease or even increase of the bias component of the combined error of complex approaches. For instance, in a simulation study for the case with three forecasts, Miller et al. (1992) showed that SA, in contrast to other approaches that learn weights from the training sample, benefits from several types of structural breaks such as location shifts. Diebold and Pauly (1987) found that structural changes in time series generally lead to decreased robustness of more complex approaches and weights that are increasingly differing from the ones that would be optimal in the evaluation sample. To decrease the out-of-sample error variance of the combined forecast in these cases, the authors proposed placing more emphasis on recent errors of the forecasts in the weight estimation.

On the other hand, weight estimates have to be treated as random variables, as for instance noted by Smith and Wallis (2009) and Claeskens et al. (2016), since the weights minimizing the error variance in the evaluation sample are unknown and thus have to be estimated from the training sample. This results in an increased combined error variance for finite samples and can cause the low empirical performance of learned weights. In contrast, simple techniques –such as SA– do not estimate weights and the errors consequently do not have a variance component while exhibiting an increased bias.

In general, SA can be expected to perform well with similar error variances of the forecasts and low or medium error correlations (Bunn, 1985; Gupta and Wilton, 1987) since the weights minimizing the combined error variance in the evaluation sample are then close to equal weights and the potential reduction of the bias component is small.

Besides this relationship, in an early theoretical work, Dickinson (1973) found that the error variance reduction resulting from forecast combination with estimated weights is often smaller than expected because of sampling issues. The author showed that the confidence intervals for weight estimates are often very broad, which indicates a high uncertainty. Winkler and Clemen (1992) additionally included correlation between the errors of different forecast mechanisms and concluded that the probability of strongly deviating weights between training and test sample is considerable, in particular when the error variances of individual forecasts are similar. Smith and Wallis (2009) showed that estimation uncertainty of the parameters can easily result in SA performing better than estimated weights when the variance-minimizing weights in the evaluation sample are close to equal weights. Elliott (2011) analyzed potential accuracy gains of OW over SA combination and identified bounds on the error covariance matrix when gains are too small to balance estimation errors. In addition, Claeskens et al. (2016) derived analytically that combinations of individually unbiased forecasts can be biased, overall leading to a higher error variance than expected.

Unstable weight estimates have also been found to be the key to the high competitiveness of SA in Monte Carlo simulations (Kang, 1986; Gupta and Wilton, 1987) and comparable results are observed with real-world data. For instance, Figlewski and Urich (1983) combined forecast for the U.S. money supply and found that model instability resulting from sampling issues prevented complex methods from performing better than SA. Kang (1986) and Clemen and Winkler (1986) examined combinations of GNP forecasts and consistently concluded that OW are too unstable to perform well while SA does not have extreme weights, leading to higher robustness and better predictive performance .

The issues with unstable weight estimates can be illustrated using the previous example in Figure 2.3. In the example, OW corresponds to a weight of 2 for one forecast (and consequently weight -1 for the other) when the error correlation approaches 1. The weight strongly declines with decreasing error correlation, down to a weight of 1.5 with error correlation of 0.9. OW is thus prone to small differences between the error characteristics estimated from the training sample and the characteristics in the evaluation sample.

Overall, the bias–variance trade-off an its facets can be identified as a main problem underlying the issues of forecast combination methods. Although empirical guidelines on how to address the trade-off in forecast combination in practice have been introduced in the literature, the trade-off is not understood theoretically and analytically. An analytical model of the error of a combined forecast separating the bias and variance components, which is required for understanding and addressing the trade-off, does not exist.

4.2 Properties of Combination Methods

In the recent literature on prediction, the bias–variance trade-off is often addressed by using estimates that are shrinked towards a common value to reduce sensitivity to training data. This approach balances the resulting increase of the bias component and the reduction of the variance of the estimates (and consequently the reduction of the variance component of the error) in a straightforward fashion. The intuition of shrinkage can easily be summarized: extremely high or low estimates are likely to contain high error and should consequently be reduced. Hence, shrinkage aims at decreasing the sensitivity of one or several weights to training data.

In forecast combination, a common direction of shrinkage is to shrink towards SA, which addresses the trade-off between the two extremes, OW and SA, and implicitly regulates the shares of the bias and variance components. Diebold and Pauly (1990) and Aiolfi and Timmermann (2005), amongst others, successfully applied shrinkage towards SA.

More formally, OW can be linearly shrinked towards SA using a shrinkage parameter $\lambda \in [0,1]$ resulting in the weight vector $\hat{w}^{\lambda}$ defined in Equation 4.1. In the course of this work, $\hat{w}_i^{\lambda}$ is used to refer to the weight of the i-th element (corresponding to the i-th forecast) of the weight vector $\hat{w}^{\lambda}$.

$$\hat{w}^{\lambda} = \lambda w^{S} + (1-\lambda)\,\hat{w}^{O} \tag{4.1}$$

Obviously, $\lambda = 0$ corresponds to OW while $\lambda = 1$ corresponds to SA. Equation 4.1 is consequently a generalization of Equations 2.2 and 2.3. As a consequence, this weight formulation is in the following theoretical analyses and discussions of forecast combination methods.

Weights are estimated from one set of forecasts and corresponding errors (the training data) and then applied to another, previously unknown set of forecasts (the evaluation data). It is commonly assumed that forecasts are unbiased and follow a multivariate normal distribution with mean zero. The errors of the k forecasts can therefore be modeled as $E \sim \mathcal{N}_k(0,\Sigma)$ in the training sample and as $\tilde{E} \sim \mathcal{N}_k(0,\tilde{\Sigma})$ in the evaluation sample with error covariance matrices $\Sigma, \tilde{\Sigma} \in \mathbb{R}^{k\times k}$. For two forecasts, the error covariance matrices and their elements, which are used in later analyses, can be defined as shown in Equation 4.2.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tilde{\Sigma} = \begin{bmatrix} \tilde{\sigma}_1^2 & \tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 \\ \tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2 \end{bmatrix} \tag{4.2}$$

When fixed weights are used to combine forecasts, the expected error variance of the combined forecast is $\mathrm{Var}\left[\left(w^{\lambda}\right)^{\top}\tilde{E}\right]=\left(w^{\lambda}\right)^{\top}\tilde{\Sigma}w^{\lambda}$. Because of the unbiasedness assumption of the forecasts, the expected error of the combined forecast is zero and the MSE is equal to the error variance of the combined forecast. Since predefined fixed weights, such as SA, are not fitted to the error characteristics of the forecasts, they seldom minimize the combined error variance. However, if weights are estimated from training data, the estimates are subject to uncertainty, which in turn influences the combined error variance. As a consequence, the estimation uncertainty has to be considered appropriately in order to address the bias–variance trade-off.

If the weights are estimated from training data, they strongly depend on the specific training data sample drawn from the population and have to be treated as random variables as a consequence. The distribution of the weight estimate random variable can be described in the form of the sampling distribution, which is introduced for OW shrinked towards SA next. Subsequently, the sampling distribution is used to determine the expected combined out-of-sample error variance of a combination.

4.2.1 Sampling Distribution of Weight Estimates

The distribution of the weight estimates can be characterized by the expected weight estimates and the covariance of the weight estimates. As shrinked weights are a linear combination of OW estimates and SA, the expectation of shrinked weights can also be expected to be a linear combination of the expectations of the two extremes. The expectation of $\hat{w}^{\lambda}$ is shown in Theorem 4.1.

Theorem 4.1 (Expectation of Shrinked Weights)**.** *The expectation of the k optimal weights estimated from a sample with error covariance matrix Σ and shrinked with $\lambda \in [0,1]$ towards equal weights is*

$$E\left[\hat{w}^{\lambda}\right]=\lambda\frac{1}{k}\vec{1}+(1-\lambda)\frac{\Sigma^{-1}\vec{1}}{\vec{1}^{\top}\Sigma^{-1}\vec{1}}$$

The expectation defined in Theorem 4.1 can be adapted to the case with two forecasts, which allows expressing the expectation using the two error variances and the error correlation (instead of the error covariance matrix), as shown in Theorem 4.2.

Theorem 4.2 (Expectation of Shrinked Bivariate Weights)**.** *The expectation of bivariate optimal weights estimated from a training sample (with error variances σ_1^2, σ_2^2*

and error correlation ρ) shrinked towards equal weights with $\lambda \in [0,1]$ is

$$E\left[\hat{w}^{\lambda}\right] = \left(\frac{\frac{\lambda}{2}\left(\sigma_1^2 - \sigma_2^2\right) + \sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \frac{\frac{\lambda}{2}\left(\sigma_2^2 - \sigma_1^2\right) + \sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \right)$$

Theorem 4.2 furthermore allows additional insight into how the shrinkage of the weights works. For both weights, the difference between the error variance is shrinked using the shrinkage factor λ. In case of $\lambda = 0$, only one error variance is included in the weight calculation (σ_2 for $\hat{w}_1^{\lambda}$ and σ_1 for $\hat{w}_2^{\lambda}$), resulting in strong weights. With increasing λ, the influence of the other error variance is increasing until both error variances have equal influence (resulting in equal weights). Consequently, shrinkage in a sense uses the error variance of one forecast to dampen the influence of the error variance of the other forecast.

While deriving the expectation of shrinked weights is straightforward, the sampling covariance of OW –and consequently of shrinked weights– is more complex and has not yet been derived. The sampling covariance of $\hat{w}^{\lambda}$ is shown in Theorem 4.3.

Theorem 4.3 (Sampling Covariance of Shrinked Weights)**.** *Defining a modified covariance matrix of the k forecasts*

$$\Sigma' = \begin{bmatrix} \Sigma'_{11} & \Sigma'_{12} \\ \left(\Sigma'_{12}\right)^{\top} & \Sigma_{k,k} \end{bmatrix}$$

where

$$\Sigma'_{11} \in \mathbb{R}^{k-1,k-1} \textit{ with } \left(\Sigma'_{11}\right)_{i,j} = \Sigma_{k,k} - \Sigma_{i,k} - \Sigma_{k,j} + \Sigma_{i,j}$$
$$\Sigma'_{12} = \left(\left[\Sigma_{k,k} - \Sigma_{k,1}\right], \ldots, \left[\Sigma_{k,k} - \Sigma_{k,k-1}\right]\right)$$

Then the sampling covariance matrix of the OW estimates (estimated from a training sample of size n and with error covariance matrix Σ) shrinked towards equal weights with $\lambda \in [0,1]$ is

$$\Omega^{\lambda} = \frac{(1-\lambda)^2}{n-k-1}\Omega^{O}\left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top}\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}\right)$$

with

$$\Omega^{O} = \begin{bmatrix} \left(\Sigma'_{11}\right)^{-1} & \Omega_{12} \\ \Omega_{12}^{\top} & \Omega_{22} \end{bmatrix}$$

$$\Omega_{12} = \left(\left[-\sum_i \left(\Sigma'_{11}\right)^{-1}_{1,i} \right], \ldots, \left[-\sum_i \left(\Sigma'_{11}\right)^{-1}_{k-1,i} \right] \right)$$
$$\Omega_{22} = \sum_{i,j} \left(\Sigma'_{11}\right)^{-1}_{i,j}$$

Three basic and intuitive properties of the sampling covariance can be derived from Theorem 4.3. First, the sampling covariance decreases with increasing shrinkage because of $(1-\lambda)^2$ in the numerator. Increasing the shrinkage factor decreases the sensitivity to different training samples. In the extreme case with $\lambda = 1$ there is no sensitivity. The decreasing sensitivity directly results in a decrease of the sampling variance. Interestingly, the relationship is not a linear but a quadratic one, halving the shrinkage factor thus quadruples the sampling covariances. Second, the sampling covariance increases with decreasing size of the training sample size (as a result of n in the denominator). Small training samples decrease the stability of the estimator, which makes extreme estimates more likely, which in turn increases the sampling variance. Third, the sampling covariance also increases with increasing number of included forecasts as more parameters have to be estimated.

The sampling covariance matrix for the general case introduced in Theorem 4.3 can again be adapted to the bivariate case, as presented in Theorem 4.4.

Theorem 4.4 (Sampling Covariance of Shrinked Bivariate Weights)**.** *The sampling covariance of bivariate optimal weights estimated from a training sample (size n, error variances σ_1^2, σ_2^2 and error correlation ρ) shrinked towards equal weights with $\lambda \in [0,1]$ is*

$$\Omega^\lambda = \frac{(1-\lambda)^2}{n-3} \frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2\right)^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

The variance of the weight estimate found in Theorem 4.4 matches the results of Winkler and Clemen (1992), who derived the variance for the bivariate case depending on the ratio of the error variances instead of the original error variances.

The introduced expectation and sampling covariance of the shrinked weights allow understanding how the weights are distributed for different training samples from the sample population. The distribution can be used as a basis for the expected combined out-of-sample error variance, which is derived next.

4.2.2 Expected Combined Out-of-Sample Error Variance

For the derivation of the error variance of the combined forecast, a reasonable and simplifying assumption is that the weights estimated from the training sample and the forecast errors in the evaluation sample have zero covariance, i.e., $\forall i,j \in \{1,\ldots,k\} : \mathrm{Cov}\left(\tilde{E}_i, \hat{w}_j^\lambda\right) = 0$. This assumption directly results from the modeling of the training and evaluation samples and the symmetrical distribution of individual forecast errors around zero. It should be noted that independence of weights estimated from the training sample and errors in the evaluation sample is not assumed in a general sense. Clearly, weight estimates typically have negative covariance with out-of-sample error variance and *absolute* error levels, and weights are similar between training and evaluation sample. The assumption is that the covariance between non-absolute error levels and weight estimates is zero because of the symmetrical distribution of errors around zero as unbiased forecasts are assumed.

Using this assumption, Theorem 4.5 introduces the combined out-of-sample error variance.

Theorem 4.5 (Expected Error Variance). *Assuming* $\forall i,j \in \{1,\ldots,k\}$: *Cov* $\left(\tilde{E}_i, \hat{w}_j^\lambda\right) = 0$*, the expected error variance resulting from combining k forecasts with error covariance matrix* $\tilde{\Sigma}$ *using weights learned from a training sample (sample size n and error covariance matrix* Σ*) is*

$$Var\left[\left(\hat{w}^\lambda\right)^\top \tilde{E}\right] = \sum_{i,j} \tilde{\Sigma}_{i,j}\Omega_{i,j}^\lambda + \sum_{i,j} \tilde{\Sigma}_{i,j} E\left[\hat{w}_i^\lambda\right] E\left[\hat{w}_j^\lambda\right]$$

The combined error variance with shrinkage factor λ in Theorem 4.5 strongly resembles the bias–variance trade-off. The first term of the combined error variance is driven by the estimation uncertainty. Predefined fixed weights, such as SA, do not vary between training samples and consequently have zero variance and covariance. In this case, the first term is zero. In contrast, the second term is the part of the error variance of the combined forecast that relates to the (in-sample) bias of a combination. The in-sample bias is in this case the difference between the combined error variance with the shrinked weights and the weights that are optimal in the evaluation sample. Hypothetically, learning weights optimal for $\tilde{\Sigma}$ would lead to the minimal value of the second term and can even be zero if forecast errors neutralize each other perfectly. Learning optimal weights however increases the first term of the equation.

As shown by Claeskens et al. (2016), a violation of the assumption used in the derivation would lead to a biased combination and an additional increase in error variance. Both effects could increase the combined error variance in comparison to Theorem 4.5, in particular for combinations with low shrinkage factors since equal weighting does not introduce the effect. Consequently, a violation of the assumption would recommend stronger shrinkage levels.

Theorem 4.6 introduces the expected combined error variance for the special case with two forecasts.

Theorem 4.6 (Expected Error Variance With Two Forecasts)**.** *In the bivariate case, where weights are estimated from a training sample (sample size n, error variances and error correlation $\sigma_1^2, \sigma_2^2, \rho$), the combined error variance in the evaluation sample (error variances and error correlation $\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \tilde{\rho}$) is*

$$
\begin{aligned}
&Var\left[\left(\hat{w}^{\lambda}\right)^{\top}\tilde{E}\right] = \frac{1}{\left(\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2\right)^2}\Bigg(\\
&\frac{(1-\lambda)^2}{n-3}\left(1-\rho^2\right)\sigma_1^2\sigma_2^2\left(\tilde{\sigma}_1^2 + 2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 + \tilde{\sigma}_2^2\right) \\
&+ \tilde{\sigma}_1^2\left(\frac{\lambda}{2}\left(\sigma_1^2 - \sigma_2^2\right) + \sigma_2^2 - \rho\sigma_1\sigma_2\right)^2 \\
&+ \tilde{\sigma}_2^2\left(\frac{\lambda}{2}\left(\sigma_2^2 - \sigma_1^2\right) + \sigma_1^2 - \rho\sigma_1\sigma_2\right)^2 \\
&+ 2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2\left(\frac{\lambda}{2}\left(\sigma_1^2 - \sigma_2^2\right) + \sigma_2^2 - \rho\sigma_1\sigma_2\right)\left(\frac{\lambda}{2}\left(\sigma_2^2 - \sigma_1^2\right) + \sigma_1^2 - \rho\sigma_1\sigma_2\right)\Bigg)
\end{aligned}
$$

Special cases of Theorem 4.6 for SA ($\lambda = 1$) and OW ($\lambda = 0$) can easily be derived as shown in Equation 4.3 and 4.4 under the assumption of identical error covariance matrices in the training and the evaluation sample.

$$
\mathrm{Var}\left[\left(w^{S}\right)^{\top}\tilde{E}\right] = \frac{1}{4}\tilde{\sigma}_1^2 + \frac{1}{4}\tilde{\sigma}_2^2 + \frac{1}{2}\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 \tag{4.3}
$$

$$
\begin{aligned}
\mathrm{Var}\left[\left(\hat{w}^{O}\right)^{\top}\tilde{E}\right] =& \frac{1}{\left(\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2\right)^2}\Bigg(\frac{1}{n-3}\left(1-\rho^2\right)\sigma_1^2\sigma_2^2\left(\tilde{\sigma}_1^2 + 2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2 + \tilde{\sigma}_2^2\right) \\
&+ \tilde{\sigma}_1^2\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)^2 + \tilde{\sigma}_2^2\left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)^2 \\
&+ 2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)\left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)\Bigg)
\end{aligned}
$$

$$=\frac{1}{n-3}\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2\right)^2}\left(\tilde{\sigma}_1^2+2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2+\tilde{\sigma}_2^2\right) \tag{4.4}$$

On the basis of the derived sampling distribution of OW estimates shrinked towards equal weights, the expected out-of-sample error variance has been quantified for the general case and for selected special cases in the formulae introduced in this section. These formulations of the expected combined error variance not only allow identifying how strong the influence of the bias and variance components are in a combination, but also enables further analysis regarding different aspects. First of all, the robustness of the combined error variance depending on the shrinkage level is analyzed in the next section by deriving thresholds regarding training sample size as well as changes in the error covariance matrix.. Subsequently, an optimal shrinkage level is derived that minimizes the expected out-of-sample error variances and different combinations are related to derive a robust level of shrinkage.

4.3 Robustness in Forecast Combination

Depending on the chosen shrinkage level, combinations are influenced by estimation uncertainty to a different extent. While combinations with very strong shrinkage are only weakly influenced, combinations with weak shrinkage are strongly affected. A main aspect influencing the estimation uncertainty is the size of the training sample used for the estimation. In order to assess the robustness of a forecast combination against small training samples in comparison to an alternative combination, the first part of this section focuses on the training sample size.

However, not only instable parameter estimates and small training sample sizes must be considered in applications in practice. Another factor potentially influencing out-of-sample performance are structural changes between the training and the evaluation sample. The previously introduced expected combined error variance is used in the second part of this section to analyze how changes in the error covariance matrix affect the combined error variance. Using these results, critical changes of individual error variances or error correlations are derived. These critical changes quantify how much individual error characteristics are allowed to change for a combination with one shrinkage factor still performing at least as good as an alternative combination.

4.3.1 Training Sample Size

The analytical formulation of the combined error variance introduced in the previous section not only depends on the error covariance matrix but also on the sample size. Although it is clear from the formulation of the sampling distribution of the weights that larger training samples decrease uncertainty, it is unknown how the sample size influences the optimal selection of a combination model and thus the robustness of a decision.

Theorem 4.5 can be used to compare the expected out-of-sample error variances when applying two different shrinkage factors λ_1 and λ_2 to decide which of the two is preferable. In general, a lower shrinkage factor (leading to weights closer to OW) is mainly beneficial in case of a large training sample since OW can be estimated more precisely. In contrast, high shrinkage factors (weights close to equal weights) can be expected to be advantageous especially for small training samples since the otherwise high variance component is strongly reduced while accepting a slight increase of the bias component. As the two components change differently for different shrinkage factors, a critical sample size for which two combinations have the same expected combined out-of-sample error variance exists in most cases. This critical sample size is presented in Theorem 4.7.

Theorem 4.7 (Minimal Training Sample Size)**.** *The expected error variances resulting from combining k forecasts with error covariance matrix $\tilde{\Sigma}$ using weights learned from a training sample (with error covariance matrix Σ) and shrinked with λ_1 and λ_2 are equal for sample size*

$$\mathring{n} = \frac{\left((1-\lambda_1)^2 - (1-\lambda_2)^2\right) \sum_{i,j} \tilde{\Sigma}_{i,j} \Omega^{O}_{i,j} \left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top} \left(\Sigma'_{11}\right)^{-1} \Sigma_{12}\right)}{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(E\left[\hat{w}_i^{\lambda_2}\right] E\left[\hat{w}_j^{\lambda_2}\right] - E\left[\hat{w}_i^{\lambda_1}\right] E\left[\hat{w}_j^{\lambda_1}\right]\right)} + k + 1$$

The term $k+1$ is trivial and indicates that the critical sample size must exceed the number of forecasts. Learning weights requires inverting the estimated covariance matrix, which is only possible if the estimated matrix is non-singular, which in turn requires at least $k+1$ observations for the covariance calculation.

The ratio term in the equation reveals two additional interesting relationships. First, $\mathring{n}$ not only depends on the difference between λ_1 and λ_2, but also on the levels of the shrinkage factors themselves. For instance much more observations are required to prefer a model with shrinkage of 20 % over one with a shrinkage of 30 % compared to preferring a 70 % shrinkage over a 80 % shrinkage. This can be directly derived from $(1-\lambda_1)^2 - (1-\lambda_2)^2$ in the numerator of the term. Reformulating to $(\lambda_1 - \lambda_2)(\lambda_1 + \lambda_2 - 2)$ illustrates that the numerator grows with increasing difference between the two shrinkage factors and is closer to zero when both shrinkage factors are close to one. Second, $\mathring{n}$ is less likely to have extreme values for higher numbers of forecasts. The denominator is clearly closely related to the differences of the bias components of the combinations with the two different shrinkage factors. OW close to equal weights are more likely for lower numbers of forecasts (because higher numbers of forecasts enable a closer fit), in turn leading to a higher probability of a low difference of the bias components. In these cases, a much larger training sample is required for the weaker shrinkage to outperform the stronger one.

Clearly, the critical sample size derived from Theorem 4.7 is not a whole number. As a consequence, rounding of the sample size is required. In most cases, rounding the critical value upwards is more appropriate since it slightly favors lower shrinkage factors, which tend to be more robust.

An interesting special case of Theorem 4.7 is the comparison with SA ($\lambda = 1$). Equation 4.5 allows determining the training sample size required to make a specific alternative shrinkage factor a reasonable choice in comparison to SA.

$$\mathring{n} = \frac{(1-\lambda)^2 \sum_{i,j} \tilde{\Sigma}_{i,j} \left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^\top \left(\Sigma'_{11}\right)^{-1} \Sigma_{12} \right) \Omega^O_{i,j}}{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(\frac{1}{k} - \mathrm{E}\left[\hat{w}^\lambda_i\right] \right) \left(\frac{1}{k} - \mathrm{E}\left[\hat{w}^\lambda_j\right] \right)} + k + 1 \tag{4.5}$$

An additional interesting special case of Equation 4.5 is presented in Theorem 4.8 where a shrinkage factor is compared to SA in the bivariate case under the assumption of unchanged error characteristics, i.e., unchanged error covariance matrix, between training and evaluation sample. This special case has the most simplifying assumptions, but however results in a compact formulation of the minimal training sample size. Despite the simplifications, it resembles practical settings closely. First, in practice, different shrinkage factors are seldom compared regarding training sample size. Instead, different shrinkage factors are compared to SA, which is always a robust alternative. Second, while there is little information available about the error characteristics underlying the training sample (since only one estimate is available), there is by definition no information available regarding the evaluation sample. The only reasonable assumption is consequently that the error characteristics do not change between the two samples.

Theorem 4.8 (Minimal Sample Size for the Bivariate Case Without Changes). *In the bivariate case, assuming error variances and correlation $\sigma_1^2, \sigma_2^2, \rho$ (unchanged between training and evaluation sample), the expected combined error variance with weights learned from the training sample and shrinked with λ is equal to the combined error* *variance with SA for training sample size*

$$\mathring{n} = (1-\lambda)^2 \left(1-\rho^2\right) \left(\frac{2\sigma_1\sigma_2}{\sigma_2^2 - \sigma_1^2}\right)^2 + 3 \tag{4.6}$$

The simplified case shows several influences on the minimal training sample size. First, as already discussed for the general case, the training sample size increases with lower levels of shrinkage. Second, the minimal training sample size decreases with increasing (absolute) correlation. As already illustrated in Section 2.2, strong correlations allow strong weights and consequently an expected substantial reduction of the bias component. Even for relatively low training samples, this effect is not completely negated by the increased variance component. Third, $\mathring{n}$ decreases for increasing differences in error variance between the forecasts. This result is also closely related to the previous discussions, as similar error variances result in a low potential reduction of the bias component, which is easily negated by the increased variance component. For $\sigma_1 = \sigma_2$, the value of the minimal training sample size even reaches infinity as SA is the optimal choice in this case.

While the first result was already clear from the previous analyses of the general case, the other results can also be expected to be transferable to an arbitrary number of forecasts. Thus, if the diagonal elements of the error covariance matrix, i.e., the error variances, are very similar, the training sample size required to outperform SA is rather large. In contrast, the sample size decreases with increasing values of the non-diagonal elements.

Overall, the introduced minimal training sample size allows comparing models with different shrinkages and determining whether the available training data is large enough for a combination to outperform an alternative one. This aspect addresses the bias–variance involved in the combination by balancing errors resulting from the estimation uncertainty and from undersensitivity to the training data. Thus, under the assumption that the population error covariance matrix is known for the training sample and, additionally, that it is identical to the one of the evaluation sample, models can be compared using the training sample size.

However, the error covariance matrices must not necessarily be identical for past and future data in practical settings. Structural changes, which are analyzed next, can result in significant differences that in turn influence the robustness of a decision.

4.3.2 Error Covariance Change

Changes of individual error variances or correlations can result in the parameter estimated from the training sample not being very well suited for the evaluation sample. A simple example is a combination where one of the forecasts was very accurate in the training sample, but, for some reason, performs substantially worse in the evaluation sample. The forecast is assigned a high weight because of its accuracy in the training sample, which however substantially increases the combined error variance when applied to the evaluation sample. In contrast to estimated weights, SA is affected far less by this change as the forecast with the increasing error variance has equal weight as the other forecasts.

The definition of the combined error variance in Theorem 4.5 uses two completely different error covariance matrices Σ in the training sample and $\tilde{\Sigma}$ in the evaluation sample. Although, in general, the two covariance matrices can change completely, changes are more likely to occur for a limited set of error variances or error correlations. Assuming that only one error variance or error correlation within the error covariance matrix changes, the influence of the change on the combined error variance can be analyzed.

To enable a compact formulation, a new definition is introduced in Equation 4.7 so that the combined error variance can be expressed as $\sum_{i,j} \tilde{\Sigma}_{i,j} \Psi^{\lambda}_{i,j}$. The elements of Ψ^{λ} in a sense indicate how strongly an element of the error covariance matrix influences the combined error variance, either by having a high weight or by being very uncertain.

$$\Psi^{\lambda} = \Omega^{\lambda} + \mathrm{E}\left[\hat{w}^{\lambda}\right] \mathrm{E}\left[\hat{w}^{\lambda}\right]^{\top} \tag{4.7}$$

In a first step, the error correlation between two forecasts p and q is assumed to change by Δ_{ρ} while all other error correlations as well as the error variances are fixed. The resulting combined error variance can be defined as shown in Theorem 4.9. Interestingly, the combined error variance can be expressed as the combined error variance without change with an additional adjustment term.

Theorem 4.9 (Combined Error Variance with Correlation Change)**.** *Assuming that the error correlation between forecasts p and q changes by Δ_{ρ} between training and evaluation sample, the combined out-of-sample error variance resulting from combining k forecasts with weights learned from a training sample (sample size n and error covariance matrix Σ) and shrinked with $\lambda \in [0, 1]$ is*

$$Var\left[\left(\hat{w}^{\lambda}\right)^{\top} \tilde{E}\right] = \sum_{i,j} \Sigma_{i,j} \Psi^{\lambda}_{i,j} + 2\Delta_{\rho} \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Psi^{\lambda}_{p,q}$$

The first term in Theorem 4.9 is the combined error variance for $\Sigma = \tilde{\Sigma}$, the second term then corrects the error covariance between forecasts p and q for the change of the covariance by Δ_ρ. Clearly, the influence of a correlation change is determined by the influence of the corresponding element of the error covariance matrix on the combined error variance, as quantified by Ψ^λ. The case of multiple correlation changes can be considered by introducing additional correction terms.

The impact of a change of the error variance of an individual forecast can be analyzed in a similar manner. Assuming a change of the error standard deviation of forecast p by Δ_σ, the combined error variance can be reformulated as presented in Theorem 4.10, where the second term corrects for the additional error variance resulting from the variance change while the third terms corrects for the covariance changes.

Theorem 4.10 (Combined Error Variance with Variance Change)**.** *Assuming that the error standard deviation of forecast p changes by Δ_σ between training and evaluation sample, the combined out-of-sample error variance resulting from combining k forecasts with weights learned from a training sample (sample size n and error covariance matrix Σ) and shrinked with $\lambda \in [0,1]$ is*

$$Var\left[\left(\hat{w}^\lambda\right)^\top \tilde{E}\right] = \sum_{i,j} \Sigma_{i,j}\Psi^\lambda_{i,j} + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j}\Psi^\lambda_{p,j} + \left(2\Delta_\sigma\sqrt{\Sigma_{p,p} + \Delta_\sigma^2}\right)\Psi^\lambda_{p,p}$$

As can be seen in Theorem 4.10, the influence of a variance change is also determined by the influence of the elements of the error covariance matrix affected by the change on the combined error variance, as quantified by Ψ^λ. The affected elements are, in contrast to the correlation change, not only the error variance itself, but also all error covariances involving the forecast.

Overall, the introduced formulae quantify how the combined error variance changes if an error correlation or error variance changes from its value in the training sample. The influence of a change depends, as can be expected, on the importance of the changed element of the error covariance matrix for the combined error variance, i.e., whether the element had a high weight or high uncertainty.

As Theorems 4.9 and 4.10 allows determining the impact of changes to the error covariance matrix, the impacts can be compared for combinations with different shrinkage levels. Since the impact depends on the value chosen for the shrinkage factor λ, combinations with different levels of shrinkage are more or less prone to structural changes. As a result of the high sensitivity to different training samples, combinations with small λ are more prone to changes. In contrast, a strong shrinkage with a high λ is more robust to changes. Consequently, a strength of change of error correlation or variance must in many cases exist, for which combinations with different shrinkage factors have equal expected combined out-of-sample error variance.

Before these critical changes are identified, an additional definition is introduced in Equation 4.8 for convenience, based upon the previously defined matrix Ψ^{λ}.

$$\Delta\Psi^{\lambda_1,\lambda_2} = \Psi^{\lambda_1} - \Psi^{\lambda_2} \tag{4.8}$$

Using this definition, Theorem 4.11 introduces the critical error correlation change, i.e., how much an error correlation is allowed to change for a decision for a combination with shrinkage λ_1 to still perform at least as good as an alternative combination with shrinkage λ_2.

Theorem 4.11 (Critical Correlation Change)**.** *The combined out-of-sample error variances resulting from combining k forecasts with weights learned from a training sample (sample size n and error covariance matrix Σ) are equal for shrinkage with λ_1 and λ_2 in case of a change of the error correlation between forecasts p and q between training and evaluation sample by*

$$\mathring{\Delta}_{\rho} = -\frac{\sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1,\lambda_2}}{2\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\Delta\Psi_{p,q}^{\lambda_1,\lambda_2}}$$

The numerator in Theorem 4.11 is equal to the first term in Theorem 4.9 with $\Delta\Psi^{\lambda_1,\lambda_2}$ instead of Ψ^{λ}. The numerator is consequently the difference in combined error variance if the error covariance matrix is unchanged. For an interpretation of the denominator, the elements in $\Delta\Psi^{\lambda_1,\lambda_2}$ can be interpreted as a quantified influence of the corresponding element of the error covariance matrix on the difference between the combined error variance of the two combinations. The denominator consequently reflects the influence of the original error covariance of forecasts p and q on the difference in combined error variance. As a result, the critical value is small (high) if the changing error covariance is responsible for a large (small) portion of the difference.

Likewise, a critical change of individual error variances can be derived for which two combinations with different shrinkage parameters have equal combined out-of-sample error variance. The critical error variance change is introduced in Theorem 4.12.

Theorem 4.12 (Critical Variance Change)**.** *The combined out-of-sample error variances resulting from combining k forecasts with weights learned from a training sample (sample size n and error covariance matrix Σ) are equal for shrinkage with λ_1 and λ_2*

in case of a change of the error standard deviation of forecast p between training and evaluation sample by

$$\mathring{\Delta}_\sigma = -\sum_i \frac{\Sigma_{p,i}}{\sqrt{\Sigma_{p,p}}} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{p,i}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}} \pm \sqrt{\left(\sum_i \frac{\Sigma_{p,i}}{\sqrt{\Sigma_{p,p}}} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{p,i}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}}\right)^2 - \sum_{i,j} \Sigma_{i,j} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{i,j}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}}} \tag{4.9}$$

All ratios in the Theorem 4.12 contain $\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}$ in the denominator. Since this term can be interpreted as the strength of the influence of the error variance of the p-th forecast on the difference in combined error variance, small changes of the error variance are critical if the influence is strong. In contrast, the critical value is large if the difference in influence is small.

Calculating the critical value in Theorem 4.12 in most cases leads to two solutions, $\mathring{\Delta}^+_\sigma$ and $\mathring{\Delta}^-_\sigma$. Hence, different cases have to be considered regarding critical changes in error standard deviation: either $0 \in \left[\mathring{\Delta}^-_\sigma, \mathring{\Delta}^+_\sigma\right]$ or $0 \notin \left[\mathring{\Delta}^-_\sigma, \mathring{\Delta}^+_\sigma\right]$. In the former case, decreases as well as increases in error variance over particular threshold values are critical. In the latter case, increases or decreases are critical between the two threshold values only, while stronger changes again lead to the original method performing better. For instance, if $\mathring{\Delta}^-_\sigma < \mathring{\Delta}^+_\sigma < 0$, only decreases between $\mathring{\Delta}^-_\sigma$ and $\mathring{\Delta}^+_\sigma$ are critical.

It should however be noted that changes in the covariance matrix cannot occur in an arbitrary fashion. Regarding changes of correlation, a change of a correlation by Δ_ρ must not necessarily result in a valid (i.e. positive definite) covariance matrix. Subtracting Δ^-_σ from the original error standard deviation can furthermore lead to negative values, in which case the change is invalid.

Unfortunately, both critical changes are rather complex and can consequently not be formulated compactly for the bivariate case. However, additionally restricting to SA and OW allows deriving a formulation, which is introduced in Theorem 4.13. Since both sampling issues due to small sample sizes and changes in the error covariance matrix can lead to increased out-of-sample error variance, the theorem additionally provides an analysis that isolates the effect of diverging sample characteristics from estimation errors. For this purpose, an additional critical change of the error correlation is provided under the assumption of an infinitely large training sample.

Theorem 4.13 (Critical Correlation Change With Two Forecasts for SA and OW). *The combined out-of-sample error variances of a SA and an OW combination of two forecasts are equal for a specific training sample (sample size n and error variances and correlation $\sigma_1^2, \sigma_2^2, \rho$) for a change of the error correlation between training and evaluation* *sample by*

$$\mathring{\Delta}_\rho = -\frac{\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2 - \frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{2\sigma_1\sigma_2\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2+\frac{1+\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)}$$

Assuming a training sample of infinite size, the critical change of error correlation is

$$\mathring{\Delta}_\rho^\infty = \rho - \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1\sigma_2}$$

Critical values for changes of an error variance can be defined in the bivariate case for SA and OW in a similar manner, as presented in Theorem 4.14. In order to isolate the influence of the changed variance from parameter instability, an additional critical value is again derived under the assumption of an infinite training sample.

Theorem 4.14 (Critical Variance Change With Two Forecasts for SA and OW)**.** *The combined out-of-sample error variances of a SA and an OW combination of two forecasts are equal given a training sample (sample size n and error variances and correlation $\sigma_1^2, \sigma_2^2, \rho$) for a change of the error standard deviation of forecast* 1 *between training and evaluation sample by*

$$\begin{aligned}\mathring{\Delta}_\sigma = & -\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\\ & \pm\left(\left(\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^2\right.\\ & \left.\quad-\frac{\left(\frac{1}{4}d^2-m\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^{\frac{1}{2}}\end{aligned}$$

where $d = \sigma_1^2 - \sigma_2^2$ and $m = \frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2$.

Assuming a training sample of infinite size, the critical change of error standard deviation is

$$\begin{aligned}\mathring{\Delta}_\sigma^\infty = & -\sigma_1 - \frac{\rho\sigma_2\left(\sigma_1^2-\sigma_2^2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\\ & \pm\sqrt{\left(\sigma_1+\frac{\rho\sigma_2\left(\sigma_1^2-\sigma_2^2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\right)^2-\frac{\left(\sigma_1^2-\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}}\end{aligned}$$

As already pointed out in the discussion of Theorem 4.12, two critical values can exist in some cases. In the case with two forecasts, the first solution is reached by changing σ_1 towards σ_2, in which case SA and OW at some point have equal combined error variance. The second solution, however, is reached when σ_1 decreases to a value approaching zero, in which case OW is also expected not to outperform SA. While the first solution is expected and conforms to insights from the literature and the discussion in Section 4.1, the second solution can be explained by comparing OW estimated from the training sample to the weights that would minimize the combined error variance in the evaluation sample. For example, assuming $\rho = \tilde{\rho} = 0.9$, $\sigma_2 = 1$ and $\sigma_1 = 0.6$ changing to $\tilde{\sigma}_1 = 0.01$. In this case, the OW estimate $\hat{w}_1^O$ is 1.643 whereas the weight that is optimal in the evaluation sample is $1.009 \approx 1$. In this example, equal weights are "closer" than the estimated weights – even though the difference between the error variances of the forecasts increased, which usually benefits OW.

Overall, the results of this section allow assessing the robustness of a decision. For this purpose, the training sample size and the maximum changes of error covariances can be determined that make one combination more beneficial that an alternative one. Thus, given a decision for a shrinkage level, the robustness can be evaluated in comparison to an alternative combination such as SA. As the robustness clearly depends on the chosen shrinkage level (stronger shrinkage in general increases robustness), the shrinkage level can be used to increase the robustness to a desired level, as is shown in the next section.

4.4 Combination Considering Bias–Variance and Structural Change

The previous section has shown how large training samples must be to make a specific shrinkage factor a reasonable choice. The derived training sample size addresses the trade-off between the bias and variance related components of the combined error variance for a specific shrinkage level. In the first part of this section, the shrinkage parameter that optimally solves the bias–variance trade-off and minimizes the expected error variance is introduced as the shrinkage can, in contrast to the sample size, be controlled directly.

Likewise, the introduced critical changes of error covariances allow determining the robustness of a decision for a specific shrinkage level. Different shrinkage levels can consequently be compared regarding their robustness against changes. In the second part of this section, the critical changes are used to define a robust shrinkage level, which allows a definable robustness against changes.

4.4.1 Optimal Bias–Variance Aware Shrinkage

The shrinkage level allows a direct trade-off between the bias and the variance component. While the variance component is reduced by stronger shrinkage, the bias component is increased as a result of the weaker fit to the data. The statistical learning theory motivates the assumption that the relationship between the two components is non-linear and that an optimal level of model flexibility, i.e., shrinkage, exists. Since Theorem 4.5 provides a closed formulation for the expected error variance of a combination with shrinkage parameter λ, the optimal shrinkage parameter $\mathring{\lambda}$, which is presented in Theorem 4.15, can be derived.

Theorem 4.15 (Optimal Shrinkage). *The expected error variance resulting from combining k forecasts with error covariance matrix $\tilde{\Sigma}$ using OW learned from a training sample (sample size n and error covariance matrix Σ) is minimized by using the shrinkage factor*

$$\mathring{\lambda} = \frac{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(\Omega_{i,j} - \frac{1}{2k} E\left[\hat{w}_i^O\right] - \frac{1}{2k} E\left[\hat{w}_j^O\right] + E\left[\hat{w}_i^O\right] E\left[\hat{w}_j^O\right] \right)}{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(\Omega_{i,j} + \left(\frac{1}{k}\right)^2 - \frac{1}{k} E\left[\hat{w}_i^O\right] - \frac{1}{k} E\left[\hat{w}_j^O\right] + E\left[\hat{w}_i^O\right] E\left[\hat{w}_j^O\right] \right)}$$

The formula of the optimal shrinkage shows two main relationships. First, $\mathring{\lambda}$ decreases with increasing number of forecasts k. The equation can be interpreted as the ratio of two weighted sums of the elements of $\tilde{\Sigma}$ with different weights in the numerator and the denominator. For these weights, $\Omega_{i,j}$ is equal in numerator and denominator, while the terms involving k clearly shrink much faster in the numerator for increasing k, overall leading to a decrease of the ratio. Second, $\mathring{\lambda}$ decreases with sample size. $\Omega_{i,j}$ is the only term involving the sample size and its value decreases for increasing n (since the uncertainty reflected by this term decreases). When keeping the other terms fixed, increasing the sample size results in a decrease of the optimal shrinkage factor since the other terms are substantially smaller in the numerator than in the denominator.

Unfortunately, the optimal shrinkage factor is rather complex in the bivariate case and no compact representation can be derived. The formula is consequently omitted for reasons of comprehensiveness since all analyses and discussion can also be based upon the general definition.

While the introduced optimal shrinkage level adjust for the estimation uncertainty by optimally balancing the bias and the variance component, it does not ensure additional robustness against structural changes. The robust shrinkage level that focuses on this aspect is introduced next.

4.4.2 Robust Shrinkage

While the critical changes derived in the previous section allow interesting insights into the robustness of forecast combination, they furthermore allow determining a robust shrinkage level. As the critical values are defined based on a comparison to an alternative shrinkage level, the robust shrinkage factor is in this context defined as the shrinkage level performing at least as good as an alternative shrinkage factor λ, for instance $\lambda = 1$, as long as certain robustness requirements are met.

More formally, a shrinkage factor that is robust against (positive or negative) changes of error correlation up to a definable value r can easily be derived using the critical values introduced in Theorems 4.11 and 4.12. Let $\Delta_\rho(\lambda_1, \lambda_2, p, q)$ denote the critical correlation change for forecasts p and q and shrinkage factors λ_1, λ_2 (Σ, k and n are omitted as parameters for reasons of simplicity). Using the previous definition, the shrinkage factor with robustness against correlation changes is the minimum shrinkage for which changes of the error correlation of all pairs $p \neq q$ up to a certain level r are uncritical, as defined in Equation 4.10.

$$\mathring{\lambda}_\rho(\lambda, r) = \min \left\{ l \mid \forall p \neq q : \left| \Delta_\rho(l, \lambda, p, q) \right| < r \right\} \tag{4.10}$$

Similarly, a shrinkage factor with robustness against changes of error variance can be introduced. With $\Delta_\sigma^+(\lambda_1, \lambda_2, p)$ and $\Delta_\sigma^-(\lambda_1, \lambda_2, p)$ as positive and negative critical changes for forecast p and shrinkage factors λ_1, λ_2, Equation 4.11 defines the robust shrinkage factor as the minimum shrinkage for which all relative changes by v are uncritical. For instance, $v = 0.1$ results in robustness against changes by ±10 %. In this case, relative changes are used since the error variances can, in contrast to the error correlations, differ by orders of magnitude, depending on the scale of the time series.

$$\mathring{\lambda}_\sigma(\lambda, v) = \min \left\{ l \mid \forall p : \begin{cases} \frac{\Delta_\sigma^+(l,\lambda,p)}{\sqrt{\Sigma_{p,p}}} > 1 - \sqrt{1+v} & \text{if } \Delta_\sigma^+(l, \lambda, p) < 0 \\ \frac{\Delta_\sigma^-(l,\lambda,p)}{\sqrt{\Sigma_{p,p}}} < \sqrt{1+v} - 1 & \text{if } \Delta_\sigma^-(l, \lambda, p) > 0 \\ \frac{\Delta_\sigma^-(l,\lambda,p)}{\sqrt{\Sigma_{p,p}}} > 1 - \sqrt{1+v} & \\ \qquad \wedge \frac{\Delta_\sigma^-(l,\lambda,p)}{\sqrt{\Sigma_{p,p}}} < \sqrt{1+v} - 1 & \text{otherwise} \end{cases} \right\} \tag{4.11}$$

The three cases in the equation are related to the different cases regarding positive and negative values of the two critical values, as previously discussed. While the last case is the standard one, the two other cases use the critical value closer to zero as critical value since the other again leads to the original method performing better.

If robustness against changes of error variances as well as error correlations is desired, the corresponding robust shrinkage factor is the maximum of the two introduced shrinkage factors, as shown in Equation 4.12.

$$\mathring{\lambda}_R(\lambda, r, v) = \max\left\{\mathring{\lambda}_\rho(\lambda, r), \mathring{\lambda}_\sigma(\lambda, v)\right\} \tag{4.12}$$

In summary, based upon the assumption that only one error correlation or error variance changes between training and evaluation sample, the impact of the change on the combined error variance has been analyzed. This influence in turn allowed deriving critical changes, which indicate maximum allowed changes for a combination to still outperform an alternative one. As critical changes can be calculated for different shrinkage levels, a shrinkage level can be chosen that satisfies predefined robustness requirements in the form of maximum changes. The resulting robust shrinkage level can be used instead of optimal shrinkage in order to minimize the combined error variance while ensuring a certain degree of robustness.

While it is clear that robustness requires a stronger shrinkage, the size of this effect, especially depending on parameters such as the number of forecast, is unknown. For this reason, the robust shrinkage level (and all previously derived results on forecast combination) are illustrated and discussed in the next section.

4.5 Illustration and Discussion

The previous sections first provided an analytical formulation of the combined error variance with a decomposition into a bias and a variance component. The combined error variance was then analyzed regarding different aspects, ranging from the critical training sample size and critical changes to the two introduced shrinkage levels. Although most of the analytical results were briefly discussed on the basis of the derived formulae, only relatively basic insights could be derived because of the complexity of the results. In this section, the analytical results are illustrated for an in-depth discussion of the results.

The combined error variance as well as the derived thresholds and results depend on a large number of parameters (k error variances, $\frac{k(k-1)}{2}$ error correlations, and the sample size). Consequently, analyses of the results as a function of individual parameters are only possible for the case with two forecasts. For the discussion of the general case, complete error covariance matrices are used.

For this purpose, the error covariances of the forecasts of the M3 Competition (Makridakis and Hibon, 2000) are used. In order to ensure a sufficient number of observations, only the 1,428 monthly time series are used. Of the 24 forecasting methods considered in the competition (see Makridakis and Hibon (2000) for a description of the individual approaches), the method *Comb S-H-D* is excluded for the error covariance calculation since it already is a combination of three forecasts. For each of the 23 forecasting methods, 18 monthly forecasts (and corresponding actuals) are available per time series, from which error covariance matrices can be estimated.

The error characteristics of the forecasts are presented in Table 4.1, where the error variances are scaled by the error variance of *Naïve2*. The error variances across time series vary substantially between forecasting models and on average (winsorized at the 5 % quantile) range from 94 % for *ForecastPro* to 120 % for *Flores/Pearce1*, relative to *Naïve2*. Some models furthermore perform relatively stable (e.g. *Dampen* with a very low standard deviation of 13.9 % of the error variance of *Naïve2*) while others vary strongly (for instance *Flores/Pearce1* with a standard deviation of 97.5 %). Forecast errors are strongly correlated with the lowest mean correlation between the errors of one model and all other models being 0.71 (between *Robust-Trend* and *Automat ANN*).

As the complete estimated of the error covariance matrix with 18 observations for 23 models is singular, non-singular sub-matrices of the original complete error covariance matrices are generated. For this purpose, the following iterative procedure is used. Forecasting models are sorted (either randomly or by average symmetric mean absolute percentage error (sMAPE), the main evaluation criterion of the M3 Competition) and the first model is always included in the error covariance sub-matrices. The next forecast is added only if the resulting error covariance matrix is still non-singular (a reciprocal condition number larger than 10^{-10} is used as a criterion). The procedure is repeated until 10 forecasts are included or all remaining forecasts would result in the covariance matrix being singular. Of the 1,428 time series, 153 time series did not result in covariance matrices of size 10. These matrices have size 9×9 (73 matrices), 8×8 (44 matrices), 7×7 (10 matrices), 6×6 (10 matrices) , 5×5 (11 matrices), 4×4 (2 matrices) and 3×3 (3 matrices).

For the discussion, primarily the error covariance matrices with random ordering of the forecasts are used. Only if the results for the forecasts ordered by performance differ substantially, these results are displayed additionally. As the case with two forecasts allows an explicit discussion of the derived formulae and thresholds, these relationships are additionally illustrated and discussed as a function of selected parameters. Furthermore, the results are compared to similar

	Error Variance		Error Correlation	
Forecast	Mean	Std. Dev.	Mean	Std. Dev.
Naïve2	1.000	0.000	0.852	0.294
Single	0.999	0.005	0.853	0.293
Holt	1.055	0.606	0.846	0.280
Dampen	0.959	0.139	0.867	0.264
Winter	1.037	0.526	0.849	0.277
B-J automatic	0.978	0.324	0.849	0.268
Autobox1	1.150	0.709	0.797	0.291
Autobox2	1.044	0.490	0.832	0.269
Autobox3	1.206	0.806	0.791	0.303
Robust-Trend	1.160	0.732	0.788	0.304
ARARMA	1.150	0.691	0.794	0.294
Automat ANN	1.182	0.704	0.792	0.310
Flores/Pearce1	1.285	0.975	0.807	0.324
Flores/Pearce2	1.093	0.547	0.833	0.288
PP-autocast	1.020	0.279	0.844	0.272
ForecastPro	0.942	0.375	0.856	0.262
Smart-Fcs	1.096	0.587	0.822	0.285
Theta-sm	1.072	0.314	0.818	0.275
Theta	0.947	0.320	0.866	0.246
RBF	1.055	0.555	0.833	0.245
ForecastX	0.954	0.252	0.860	0.267
AAM1	1.077	0.658	0.818	0.289
AAM2	1.124	0.712	0.803	0.297

Table 4.1: Error characteristics of the forecasts of the M3 Competition. The winsorized mean and standard deviation of error variances of the forecasts (scaled by the error variance of Naïve2) vary strongly. Mean and standard deviation of the error correlation between a model and all other models indicate correlations that are on average above 0.8 and a standard deviation of correlations around 0.28.

results from the literature for an additional validation of the derived formulae and thresholds.

First, the bias–variance trade-off in forecast combination is discusses as it motivates the derived minimal training sample size and the optimal shrinkage factor, which are discussed subsequently. In order to illustrate the importance of structural changes, the impact of a change of error covariance is analyzed next, followed by a discussion of the robust shrinkage factor.

4.5.1 Bias–Variance Trade-Off

As the derived formulation of the combined error variance and the introduced bias and variance components are the basis for most of the other results, the bias–variance trade-off is illustrated in a first step. The statistical learning theory and the theoretical analyses in the previous sections indicate that the bias (variance) component should increase (decrease) with increasing shrinkage. As this results in a U-shape of the combined error variance, an optimal shrinkage level can be expected. This shrinkage level is likely to depend on the training sample size and the number of forecasts included in the combination.

Figure 4.1 presents the two terms of the combined error variance (Theorem 4.5) and their sum for different shrinkage factors, depending on the numbers of forecasts k and the sample size n. For the analysis, the error covariance matrices of the M3 Competition are used under the assumption that the error covariance matrices do not change between training and evaluation sample (i.e. $\Sigma = \tilde{\Sigma}$). The values of the components are scaled by the MSE (i.e. sum of the components) for $k = 2$ and $\lambda = 1$ and the median of the components and their sum across the time series is displayed.

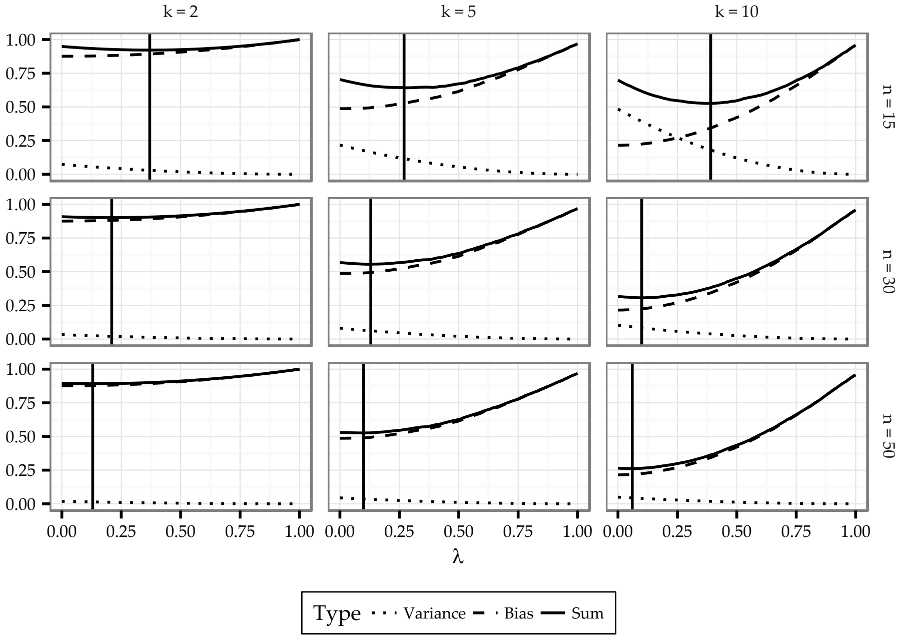


Figure 4.1: Median scaled values of the components of the MSE depending on the shrinkage factor λ. The shrinkage clearly influences the bias and variance components in different ways, depending on the number of forecasts k and the size of the training sample n. The optimal shrinkage factor minimizing the sum of both components (vertical line) decreases with sample size and with the number of forecasts as long as the sample size suffices.

The figure strongly resembles the illustrative example of the bias–variance trade-off in Figure 2.4. For the combined error variance, the trade-off (and consequently the U-shape) is stronger for higher numbers of forecasts where the influence of the variance component on the sum is substantially stronger than for fewer forecasts. Increasing sample size leads to a decrease of the variance component, easing the trade-off and allowing lower shrinkage. Overall, the optimal shrinkage factor (indicated by a vertical line) decreases with sample size and also decreases for higher numbers of forecasts as long as the sample size is not too small in relation to the number of forecasts. While the former effect is a clear result of the estimation uncertainty, the latter effect can be explained by the achievable bias reduction that is substantially larger for higher numbers of forecasts (allowing a better fit), and that outweighs the variance increase.

Overall, the bias–variance trade-off in forecast combination clearly exists and the observed relationships match the expectations, which can be derived from the statistical learning theory and the theoretical analyses in this work. Most importantly, the influence of the training sample size and the shrinkage level is illustrated, which are the focus of the next analyses.

4.5.2 Training Sample Size

The training sample size influences the uncertainty in the weight estimation and consequently the variance component of the error of the combined forecast. As discussed using the analytical results, the training sample size required for two combinations with different shrinkage to perform equal can be expected to be large for a high difference between the shrinkage levels (because of the large difference regarding the influence of the uncertainty) and for combinations of few forecasts (because of the weaker reduction of the bias component).

For an illustration of the impact of the training sample size, the empirical cumulative relative frequencies of critical sample sizes for the error covariance matrices of the M3 Competition are presented in Figure 4.2 for different shrinkage factors λ_2 when comparing with $\lambda_1 = 1$. It is again assumed that error covariance matrices are identical for training and evaluation sample ($\Sigma = \tilde{\Sigma}$). The outer left plot depicts the critical sample sizes $\mathring{n}$ required to prefer OW ($\lambda_2 = 0$) over SA ($\lambda_1 = 1$). The other plots present critical sample sizes required to outperform SA with different shrinkage levels between 0.25 and 0.75.

The two expected relationships can be identified in the figure. First, critical sample sizes decrease with increasing shrinkage level λ_2 (and thus smaller difference of shrinkages). This effect is a result of the bias–variance trade-off where a high shrinkage still offers a substantial reduction of the bias component while the increase of the variance component is relatively small. Second, smaller training samples suffice for higher numbers of forecasts to outperform SA. Based upon the earlier theoretical analysis of the equation of the critical sample size, this effect can be attributed to the fact that the bias component decreases with the number of forecasts, which especially benefits the combination with weaker shrinkage. Because of the reduced bias component with larger numbers of forecasts, smaller

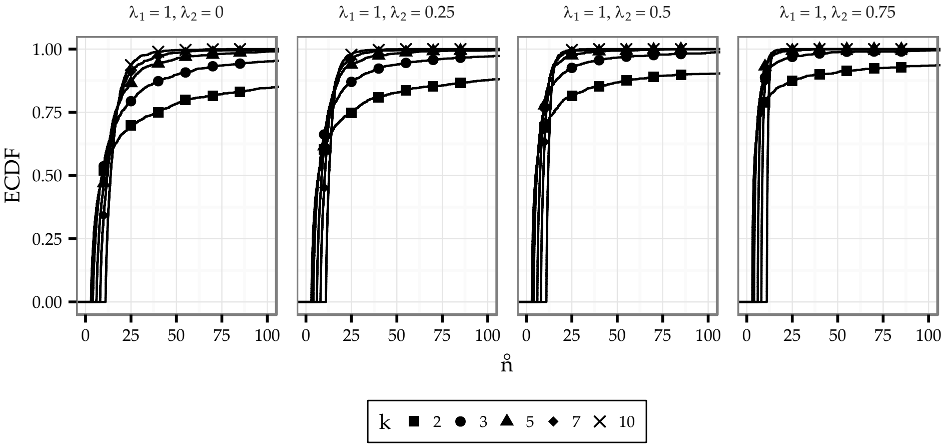


Figure 4.2: Empirical cumulative relative frequencies of critical sample sizes $\mathring{n}$ for different shrinkage parameters λ_2 and $\lambda_1 = 1$, depending on the number of forecasts k. Higher numbers of forecasts typically require smaller training samples to prefer the weaker shrinkage. While a sample size of 25 suffices for $\lambda_2 = 0.5$ in 80 % of all cases even for $k = 2$, $\lambda_2 = 0$ requires larger training samples more frequently.

training samples are required to reduce the variance component to a level where the overall combined error variance is smaller than for lower shrinkage.

For the bivariate case, the critical size of the training sample $\mathring{n}$ is depicted in Figure 4.3 as a function of ρ and for different shrinkage levels and values of σ_2 between 1.2 and 2. The value of σ_1 is without loss of generality fixed to 1, which can always be achieved by scaling.

As already noted in the general case, the critical training sample sizes decreases with stronger shrinkage. However, the bivariate case illustrates two additional effects.

First, the figure shows that $\mathring{n}$ increases with decreasing difference between $\sigma_1 = 1$ and σ_2. Consequently, the smaller the difference between the error variances of the individual forecasts, the larger the training sample required to put a combination in favor of SA. This result is consistent with the literature and intuitively clear. If error variances are very similar, OW estimates are close to equal weights, which makes the potential bias component reduction effect very small whereas the variance component still negatively affects the combined error variance.

Second, for fixed σ_2, the higher the absolute value of ρ, the smaller the required training sample. Choosing a combination other than SA always requires the largest training samples in case of uncorrelated errors and $\mathring{n}$ then decreases with increasing correlation. This relationship can, for positive error correlations,

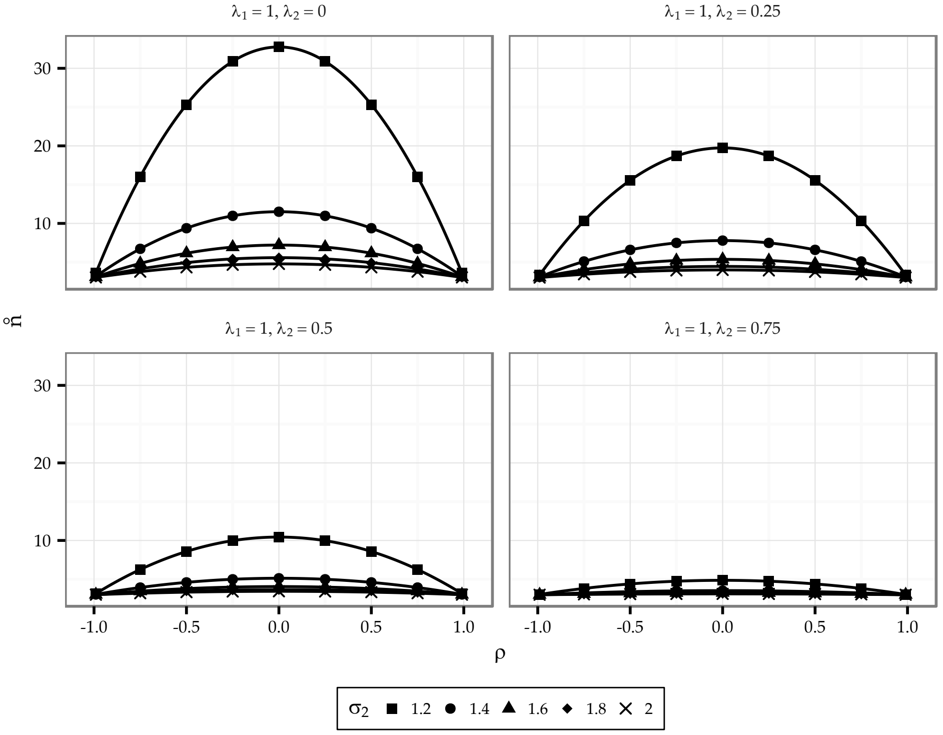


Figure 4.3: Critical training sample size $\mathring{n}$ to favor a combination with shrinkage λ_2 over SA. Cases with σ_2 close to $\sigma_1 = 1$ require larger training samples. Higher shrinkage levels and strong correlation of forecast errors decreases the required size of the training sample.

be explained on the basis of the strong OW values, as already shown in the introductory example in Figure 2.3. Strong positive error correlations can result in OW over 1 for one forecast and a negative weight for the other. The strong correlation allows strong weighting and a substantial reduction of the bias component of the combined error variance. Since extreme weights are in this case optimal, the estimated weights are likely to perform better than SA, even when accounting for estimation errors. The relationship between $\mathring{n}$ and ρ for negative values of ρ is in contrast less obvious since the error variance with SA also decreases with ρ decreasing from 0 to -1. However, as also illustrated in the introductory example in Figure 2.3, OW only differ slightly between different values of ρ as long as the correlation is strongly negative. This makes OW more robust against estimation errors of ρ. For instance, OW for $\sigma_1 = 1, \sigma_2 = 2$ results in $w_1^O = 0.714$ for $\rho = -0.5$ and $w_1^O = 0.667$ for $\rho = -1$, a difference of only around 7 %. Similarly, estimation errors of σ_1 or σ_2 do not lead to large differences in estimated weights.

These results for the case with two forecasts can be expected to generalize well to higher numbers of forecasts. OW are always close to equal weights for very similar error variances, which requires a larger training sample size to justify estimating OW independently of the number of forecasts. Likewise, correlations (i.e. non-diagonal elements of the error covariance matrix) close to zero require larger training sample sizes as the potential reduction of the bias component is smaller in this case whereas the variance components remains largely unchanged.

The training sample size threshold for SA and OW introduced in Theorem 4.8 can also easily be used to derive decision boundaries regarding error variances and correlation, given training sample size. These decision boundaries can be compared to the recommendations of Schmittlein et al. (1990) for validation. Based on the MSE of combinations with SA and OW observed in Monte Carlo simulations for different combinations of n, σ_2, and ρ, Schmittlein et al. analyzed when to use which combination method (OW with independence assumption was additionally included, which is omitted here for a direct comparison of SA and OW). While σ_1 was fixed to 1, 20 values for σ_2 and 19 values for ρ were evaluated for training sample sizes of 10, 25, 50, and 100. The results for instance suggest using SA for training samples with 10 or less error observations when error standard deviations of both forecasts differ by at most 0.2 and error correlations ρ are between -0.6 and 0.6.

Figure 4.4 displays the results of the reproduced simulation experiment, however with 10,000 instead of the 100 runs in the original experiment for more accurate results. The four plots show whether SA or OW leads to lower mean MSE across runs per parameter combination. Filled circles indicate parameter combinations where SA outperformed OW while empty circles indicate a recommendation for OW instead of SA. It should be noted that the plots differ slightly from the plots in Schmittlein et al. (1990) since results for equal error variances (i.e. $\sigma_2 = 1$) were apparently omitted there. The analytical decision boundary is indicated by a solid line.

Comparing the results in the upper-left plot with the result from Schmittlein et al. (1990) to choose SA when $|\rho| < 0.6$ and $\sigma_2 > 1.2$ for $n \leq 10$ reveals the congruence of the results. The decision boundary separates both regions precisely without mis-classifications, even when the parameter combinations are close to the boundary.

For larger sample sizes, for instance 25, the plots in Schmittlein et al. (1990) suggest using SA when error standard deviations differ by at most 0.1 and the error correlation is between -0.4 and 0.4. Comparing this recommendation with the upper-right plot in Figure 4.4 shows that this recommendation is also well covered by the boundary, which again separates both regions precisely. The decision boundaries for training sample sizes 50 and 100 at the bottom of the figure

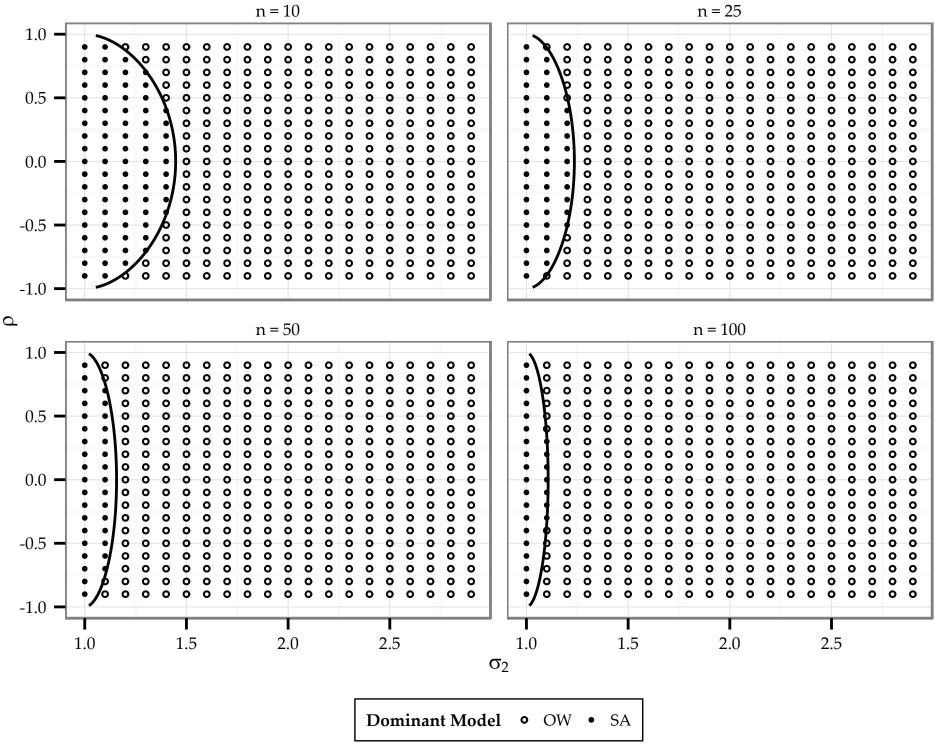


Figure 4.4: Decision boundaries (OW vs. SA) depending on σ_2, ρ, and on the training sample size n. The plots also show which model leads to lower mean MSE in a Monte Carlo experiment reproduced from Schmittlein et al. (1990). The analytical decision boundaries separate the outcomes of the simulations perfectly.

also separate the cases for and against SA precisely.

Overall, the derived critical sample size matches results from the literature well and allow a decision between two different shrinkage levels for a combination, given the size of the available training data. The optimal shrinkage level, which was introduced based on the combined error variance, however allows an even more flexible model selection.

4.5.3 Optimal Bias–Variance Aware Shrinkage

The previous discussion of the critical training sample sizes showed that relatively small training samples are sufficient to prefer a lower shrinkage level in many cases. The optimal shrinkage level, which balances the bias and the variance component of the combined error variance in an optimal way, uses this effect for a more flexible model selection.

Again assuming that error covariance matrices do not change between training and evaluation sample, the optimal shrinkage parameter from Theorem 4.15 can be calculated for the error covariance matrices of the M3 Competition. As $\mathring{\lambda}$ depends on the sample size and on the number of forecasts, the optimal shrinkage $\mathring{\lambda}$ is displayed for different values of k and n in Figure 4.5. In each plot, the solid line indicates the median of the optimal shrinkage parameters. The dark gray (light gray) areas contain 50 % (90 %) of all optimal shrinkage parameter values across the different error covariance matrices.

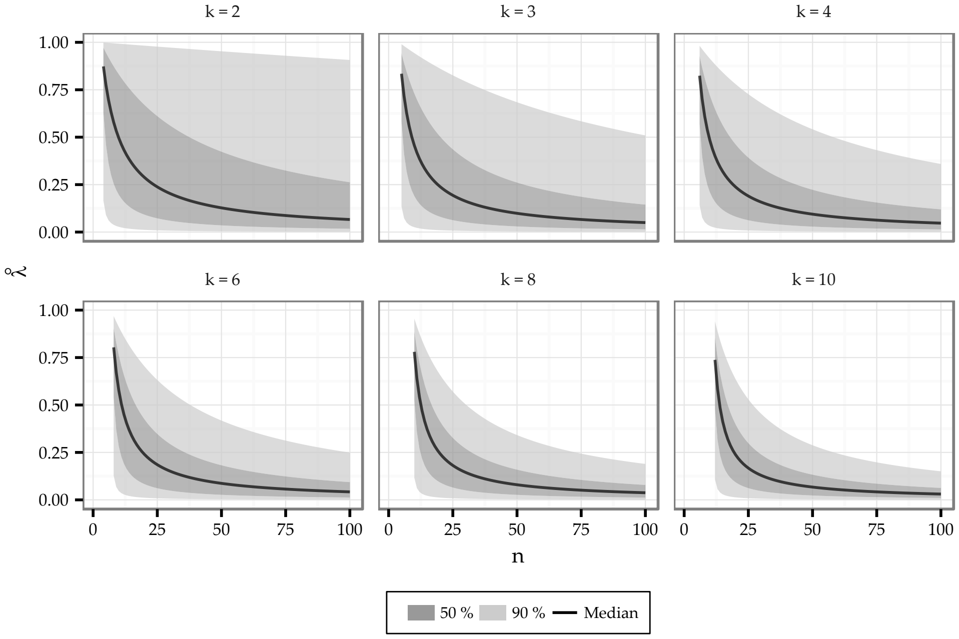


Figure 4.5: Median optimal shrinkage $\mathring{\lambda}$ (solid line) for different numbers of forecasts k and training sample sizes n. Dark gray areas contain 50 % of optimal shrinkage parameters per sample size. The optimal shrinkage parameters decrease with n. Furthermore, the uncertainty in the optimal shrinkage parameter decreases with k.

As previously derived from the analytical equation, the optimal shrinkage parameter decreases with increasing sample size for all numbers of forecasts. Starting from $\mathring{\lambda}$ exceeding 0.75 for sample size $k+1$, $\mathring{\lambda}$ decreases with a steep descent for slightly larger sample sizes. The decrease is declining with sample size and $\mathring{\lambda}$ decreases only slightly for sample sizes over 25.

Increasing the number of forecasts clearly reduces the uncertainty in optimal shrinkage parameters. While the 90 % interval nearly spans the complete spectrum of shrinkage parameters for two forecasts even for large samples, optimal shrinkage parameters are concentrated in smaller intervals for larger numbers of forecasts. In particular, the uncertainty decreases when adding additional forecasts to a combination. However, adding forecasts to a combination with five or more forecasts only leads to very small differences in $\mathring{\lambda}$.

Figure 4.6 additionally displays the optimal shrinkage level for two forecasts with $\sigma_1 = 1$ and different values of ρ, σ_2 and n. Optimal shrinkage levels increase with decreasing difference between σ_1 and σ_2 and decrease with increasing sample size. Furthermore, the optimal shrinkage decreases with increasing (absolute) correlation. Comparing the figure to the plot of the minimal training sample size (Figure 4.3) reveals an interesting similarity. Optimal shrinkage levels are low if minimal training sample sizes are also low and high if the minimal training sample size is large. This similarity is caused by both minimal training size and optimal shrinkage resulting from the bias–variance trade-off. In cases where the weights, which are optimal in the evaluation sample are far away from equal weights (e.g. for strong correlation), estimation errors are less likely to be an issue, consequently a smaller training sample or weaker shrinkage is sufficient. In contrast, when the optimal weights in the evaluation sample are close to equal weights, a large training sample is required to keep the variance component of the combined error variance low and a strong shrinkage is required.

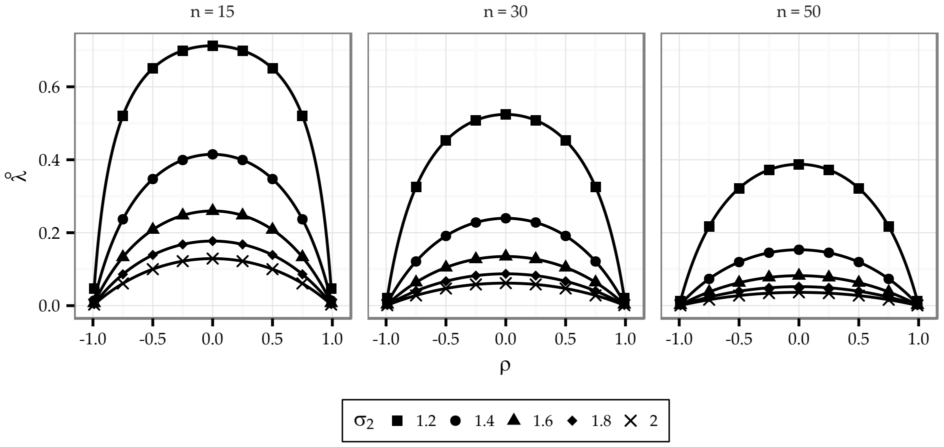


Figure 4.6: Optimal shrinkage in case of two forecasts, depending on the error correlation ρ. Results are displayed for different training sample sizes n and values of σ_2. The optimal shrinkage level decreases with training sample sizes and difference between error variances. Furthermore, stronger (absolute) correlation requires lower shrinkage.

As argued for the minimal training sample size, these basic relationships can be expected to be directly transferable to more than two forecasts

After the optimal shrinkage parameters have been analyzed, the impact of the shrinkage on the combined error variance (scaled by the corresponding combined error variance with $k = 2$) is displayed in Figure 4.7 for different numbers of forecasts and sample sizes (15, 30 and 50 as examples). The solid line again indicates the median while the dark gray (light gray) area contains 50 % (75 %) of the observations per shrinkage level. The combined error variances clearly decrease with increasing number of forecasts. The strength of the decrease increases with the size of the training sample.

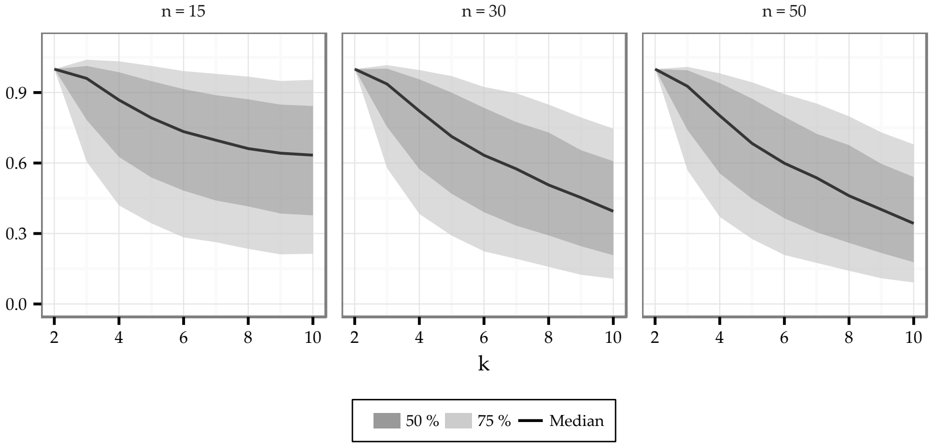


Figure 4.7: Combined error variance (scaled by the combined error variance with $k = 2$) for different sample sizes n and number of forecasts k when applying the optimal shrinkage factor. The relative combined error variance decreases with the number of forecasts and the training sample size.

The negative dependency of the expected error variance (and consequently the MSE) on the sample size follows intuitively from the decreasing variance component. The strong and monotonous effect of the number of forecasts results from the continuously decreasing bias component and would basically recommend including as many forecasts as possible since the expected error variance decreases with each additional forecast.

The result that all available forecasts should be combined with appropriately shrinked weights (as long as the training sample is not too small) to minimize the error variance clearly conflicts with existing empirical findings and guidelines, which recommend using a limited set of forecasts. Although the instability of covariance estimates, and consequently weight estimates, is accounted for, the theoretical results deviate from practical guidelines. However, the theoretical analyses have up to now only considered the uncertainty resulting from small training samples. Differences of the error covariance matrices of the training and the evaluation sample are an additional source of uncertainty. The next part of the discussion focuses on the impact of changes and on robust shrinkage.

4.5.4 Error Covariance Change

The introduced robust shrinkage factor promises a certain degree of robustness against changes of elements of the error covariance matrix. While the theoretical analyses clearly showed that changes influence the combined error variance and, consequently, the choice of the shrinkage factor, the strength of the effect is unclear. Thus, before the robust shrinkage factor is discussed, the importance of robustness is illustrated.

Since changes of error correlations do not always result in a positive definite error covariance matrix, changes in error variance are used for the illustration. In Figure 4.8, the impact of changes of the error variance of a forecast p for $n = 50$ and optimal shrinkage is displayed using the error covariance matrices from the M3 Competition. For this purpose, the expected error variance scaled by the one obtained upon no change is plotted for changes of the error variance of forecast p by -100 % to $+100$ %. In each row, the number of forecasts is increased from two forecasts in the first row to five forecasts in the last row. The forecasts are ordered by difference between the weights shrinked with optimal shrinkage and equal weights in decreasing order from left to right. Consequently, the shrinked weight of $p = 1$ has the largest positive difference to equal weights. While the solid line indicates the median, the dark gray (light gray) areas contain 50 % (75 %) of all scaled error variances per value of the change.

The median scaled combined error variance is in all cases equal to one for no change. Depending on the number of forecast and on the forecast for which the error variance changes, the combined error variance then changes more or less strongly for increasing changes.

For two forecasts (first row, $k = 2$), the influence on the combined error variance is –in comparison– relatively low. The combined error variance only slowly increases with increasing changes. For the forecast with the stronger positive weight ($p = 1$), increases in error variance can negatively influence the combination whereas decreases in error variance can even be beneficial. The reverse is true for the other forecast.

For $k = 3$, only changes of the error variance of the forecast with medium weight ($p = 2$) result in small changes of the combined error variance. For $p = 1$ ($p = 3$), increases (decreases) in error variance quickly increase the combined error variance while decreases (increases) tend to be less influential up to a certain level.

This relationship is even stronger for higher numbers of forecasts, such as $k = 4$ and $k = 5$. In these cases, changes affecting the forecasts with weight strongly deviating from equal weights (last and first forecast) immediately result in strong

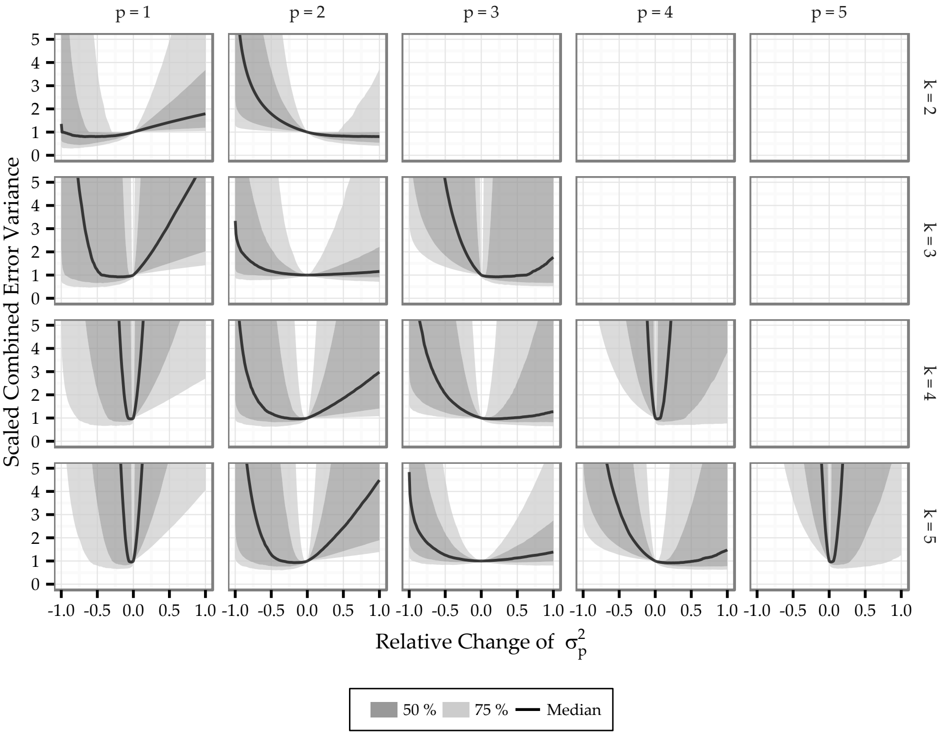


Figure 4.8: Expected combined error variance (relative to the error variance without a change) for different numbers of forecasts *k* and optimal shrinkage after a change of the error variance of forecast *p*. The forecasts are ordered by difference to equal weights in descending order. For $k = 2$, the impact of changes is relatively small. For $k = 3$, changes for the two forecasts with the highest weights can be an issue. Starting from $k = 4$, small changes can result in extreme increases in combined error variance for some forecasts.

increases in combined error variance ranging up to increases by 10 times and more.

Overall, even small changes of one error variance can have a substantial impact on the performance of the combined forecast and negate all potential benefits of a combination. The example demonstrates that in particular higher numbers of forecasts require stronger shrinkage for robustness since small changes in the error covariance matrix can otherwise result in strongly increased combined error variance that easily exceeds the error variance of a SA combination. The influence of changes can be reduced by using stronger shrinkage, which results in less extreme weights for individual forecasts. The robust shrinkage factor, which is motivated by this idea, is illustrated and discussed in the next section.

Beforehand, the case with two forecasts, restricted to SA and OW, is again used to discuss the influence of changes of error correlation and variance in greater detail.

For this case, the critical change of the error correlation $\mathring{\Delta}_\rho^\infty$ in case of two forecasts, as presented in Theorem 4.13, is depicted in Figure 4.9 as a function of ρ for different values of $\sigma_2 = \tilde{\sigma}_2$ ($\sigma_1 = \tilde{\sigma}_1$ is kept fixed at 1). The figure shows five curves for different values of σ_2 between 1.2 and 2. It is clear from the figure that only decreases of the error correlation are critical while the stronger the (positive) error correlation, the smaller the critical change. Furthermore, different values of σ_2 lead to critical changes differing by a constant factor, which is consequently independent of ρ. It can overall be stated that OW is least robust for ρ close to 1 and values of σ_2 close to $\sigma_1 = 1$.

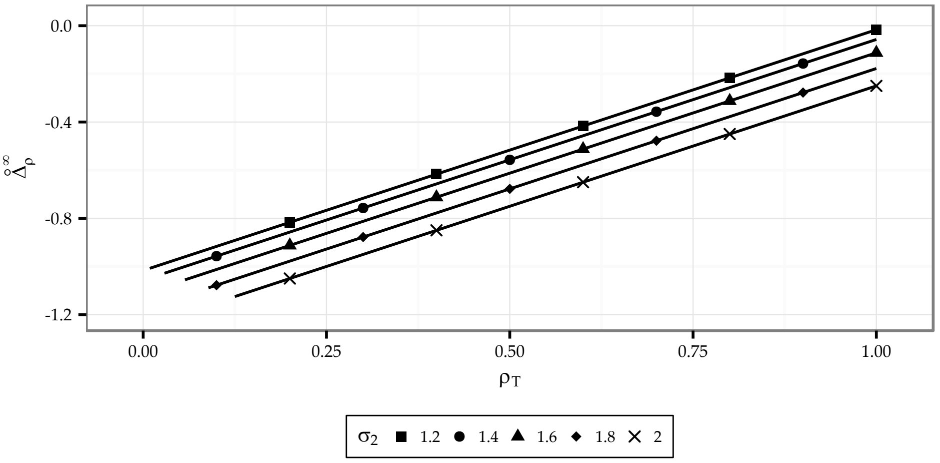


Figure 4.9: Thresholds regarding changes of error correlation in the bivariate case with SA and OW as a function of the error correlation ρ and for different values of σ_2. Forecasts with high error correlation are less robust to changes than forecasts with weak error correlation. Larger differences in accuracy between forecasts (high σ_2) increase robustness.

Both dependencies are intuitive. Strong positive error correlations can lead to extreme weight estimates with OW. In these cases, even small decreases of the error correlation lead to equal weights being closer to the weights that minimize the combined error variance in the evaluation sample. The increased robustness with increasing difference of error variances can be explained in terms of the bias–variance trade-off. In these cases, OW can decrease the bias component much more than SA. High changes of ρ are consequently required to negate this effect.

The critical change of the error standard deviation $\mathring{\Delta}_\sigma^\infty$ for two forecasts introduced in Theorem 4.14 is depicted in Figure 4.10 as a function of σ_2 between 1 and 2 (values below $\sigma_1 = 1$ are symmetrical). $\sigma_1 = \tilde{\sigma}_1$ is again fixed to 1 and the error correlation is assumed to be identical in the training and the evaluation sample, i.e., $\rho = \tilde{\rho}$. The figure shows four curves for different values of ρ between -0.99 and 0.99.

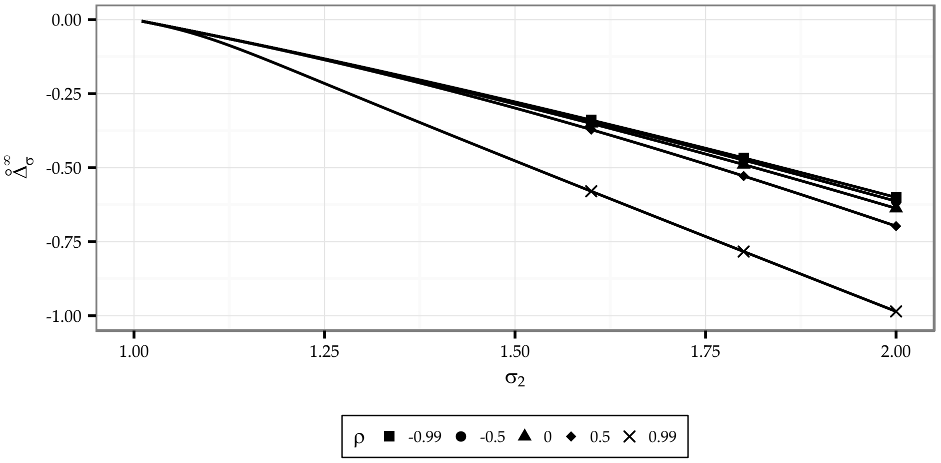


Figure 4.10: Thresholds regarding changes of error variance in the bivariate case with SA and OW as a function of the error standard deviation σ_2. Results are presented for different values of the error correlation ρ. The thresholds most strongly depend on σ_2 and thus on the difference of the accuracy of the forecasts. The error correlation only substantially influences the thresholds for strong positive correlation.

The critical change of error standard deviation increases with increasing difference between σ_1 and σ_2. The threshold values are small for very similar error variances and continuously grow in (absolute) value for increasing difference. Clearly, changes must negate a large part of the initial difference of the error variances of the two forecasts for SA to outperform the OW combination. The most pronounced case is for strong positive error correlation, where the required changes are largest. This is again due to the strong potential reduction of the bias component by using OW, which requires even larger changes to make SA perform better than OW.

The discussion of the critical changes for two forecasts can again be directly transferred to combinations of more forecasts. If the error variances are very similar, small changes of one error variance can result in SA outperforming a combination with learned weights (with and without shrinkage) if the changes occur in a disadvantageous way. If a correlation (i.e. a non-diagonal element of the error covariance matrix) changes, SA can also outperform learned weights, especially if the covariance element has –in comparison– a relatively high value.

So far, for reasons of complexity, the analytical derivations as well as the discussion were limited to a change of either error variance or error correlation, while assuming the other remained constant. However, the formula for the combined error variance introduced in Theorem 4.6 also allows a numeric analysis of a combination depending on multiple criteria. To provide further insights, numerical

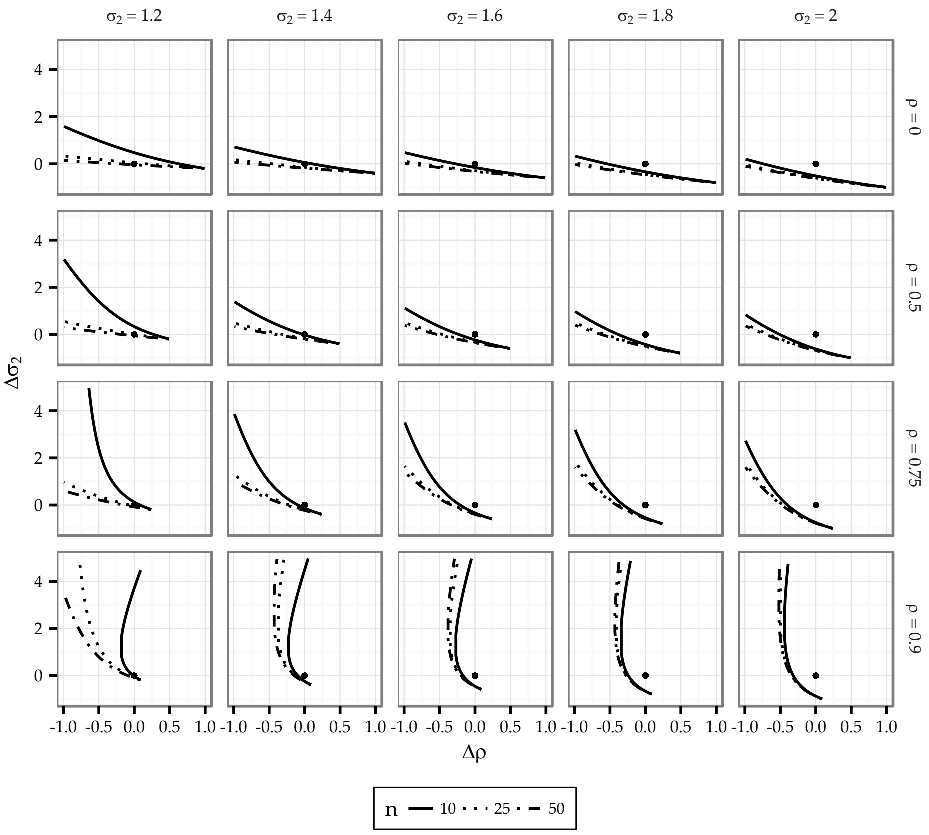


Figure 4.11: Numerical thresholds for changes of error variance as well as error correlation for different initial error characteristics ρ, σ_2 and training sample sizes. For reference, the dot in each plot marks no change of both characteristics. SA is beneficial for combinations on the left of the boundaries. Critical changes of ρ and σ_2 are dependent, SA can be beneficial as a result of combinations of changes of ρ and σ_2.

results for changes of error correlation and error variance in the bivariate case with SA and OW are presented in Figure 4.11. While $\sigma_1 = \tilde{\sigma}_1 = 1$ is again constant, combinations of changes of the error correlation and the error variance of σ_2 that result in SA and OW having equal combined error variance are displayed, depending on the initial values of σ_2 and ρ as well as on the size of the training sample n.

For combinations of changes of ρ and σ_2 on the left side of the threshold lines, SA is advantageous. Taking $\rho = 0, \sigma_2 = 2$ as an example (the plot on the upper-right of the figure), SA can be expected to perform better if either ρ decreases very strongly or σ_2 decreases substantially. Alternatively, a combination of smaller decreases of both ρ and σ_2 can result in SA being superior to OW.

The thresholds in Figure 4.11 at large confirm the previous discussions. However, two additional interesting results can be seen in the figure. First, issues with small sample sizes can be noted in some cases, especially for $\sigma_2 = 1.2$ and $n = 10$, where the no-change point is left of the threshold line. This indicates that changes in favor of OW are required to make OW a reasonable choice. Second, as previously discussed based on the analytical formula, two threshold values (one positive, one negative) for σ_2 exist for some initial error characteristics. For instance, for $\sigma_2 = 1.4$ and $\rho = 0.9$, a small negative change or a very large positive change of σ_2 with fixed ρ can result in SA performing better.

Besides the analysis and discussion of the thresholds for changes of the error characteristics, a comparison to results from the literature can provide validation and additional insights. For this purpose, the thresholds are applied to the case study of Miller et al. (1992), who analyzed the impact of breaks in error time series on different combination methods. Their simulation experiment are reproduced with three forecast error time series of length 100, which are randomly generated and combined with different combination methods. For direct comparability, only SA and OW are used. The experiments start at time interval 8, where OW are first estimated on all past observations. Then, the size of the training sample is incremented and revised weights are estimated, again on all past observations. At time 30, one characteristic of the errors (either the error variance of forecast 1, σ_1^2, or the error correlation $\rho_{1,2}$ between forecasts 1 and 2) increases or decreases from its initial value, depending on the different treatments. The initial error characteristics as well as the changes for the treatments are defined for two different sets of initial parameters, as shown in Table 4.2.

		Initial Error Characteristics							
Set	Treatment	σ_1^2	σ_2^2	σ_3^2	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{2,3}$	Change	Crit. Change
1	Variance Increase	1	0.9	1.1	0.8	0.6	0.7	$\Delta\sigma_1^2 = \quad 0.7$	∞ *
	Variance Decrease	1.7			0.8			$\boldsymbol{\Delta\sigma_1^2 = -0.7}$	**−0.295**
	Correlation Increase	1			0.4			$\Delta\rho_{1,2} = \quad 0.4$	−0.013 *
	Correlation Decrease	1			0.8			$\Delta\rho_{1,2} = -0.4$	0.514 *
2	Variance Increase	1	0.7	1.4	0.8	0.6	0.7	$\Delta\sigma_1^2 = \quad 0.8$	−0.270
	Variance Decrease	1.8			0.8			$\Delta\sigma_1^2 = -0.8$	−0.847
	Correlation Increase	1			0.4			$\boldsymbol{\Delta\rho_{1,2} = \quad 0.4}$	**0.393**
	Correlation Decrease	1			0.8			$\boldsymbol{\Delta\rho_{1,2} = -0.4}$	**−0.289**

Table 4.2: Treatments regarding the initial values of the error characteristics and their changed values after the structural break in the experiments adapted from Miller et al. (1992). The outer right column displays the analytically derived critical change (for $n = 29$) of the changing error correlation or variance. Cases where the actual change exceeds the critical change are printed in bold. A star indicates that SA is already superior to OW before the structural break. Overall, three cases of changing characteristics can be expected to be critical for the OW combination.

Using the initial characteristics and $n = 29$, the critical changes of σ_1^2 and $\rho_{1,2}$, for which OW and SA can be expected to perform similarly after the structural break, can be calculated using Theorems 4.13 and 4.14. In most cases, the threshold indicates how much the error variance or correlation is allowed to change for OW to outperform SA even after the structural change. However, in some cases (for instance the correlation decrease treatment for set 1), SA already performs better than OW. Consequently, the threshold then indicates how much error variance or correlation would have to change to make OW perform at least as good as SA after the structural break. Critical values are in these cases marked by a star. Changed characteristics exceeding the critical value are printed bold. Overall, of the eight different treatments, three can be expected to be critical changes. In the two critical cases for set 2, the changes are relatively close to the critical value, the performance of SA and OW can thus be expected to be similar after the structural break.

The average root mean squared error (RMSE) outcomes of the simulation experiments over 50,000 runs (instead of the 3,000 runs in the original experiment for higher stability of the result) are shown per time interval in Figure 4.12. For visualization purposes, the curves are slightly smoothed with smoothing splines, as done by Miller et al. (1992). It is important to note that OW continuously learns updated weights with each (seen) observation, i.e., the training period increases over time and the weights are adjusted continuously after the structural break, which explains the decreasing RMSE of OW after the structural break.

Results indicate that the derived thresholds properly differentiate between uncritical and critical cases of structural breaks. All cases where the new value of the changed characteristic is far away from the critical value are indeed uncritical. On the other hand, for the variance decrease treatment for set 1, SA is clearly superior after the structural break while it was outperformed by OW before. In the two treatments with critical changes for set 2, the RMSE values after the structural break are very close, albeit confirm the critical value and its implication. The derived analytical thresholds consequently match the results of the reproduced simulation experiment very well.

Overall, the results confirm the derived thresholds. The discussion furthermore shows that especially combinations involving many forecasts are very prone to changes of error characteristics. Even small changes can result in strong increases of the combined error variance on unknown data. As using OW estimates shrinked towards equal weights with the introduced optimal shrinkage level only considers estimation uncertainty, a higher degree of shrinkage is required for an increased robustness against changes.

4.5.5 Robust Shrinkage

A stronger shrinkage towards equal weights is required for robustness against changes since SA, as an extreme approach, is completely independent of the training sample and can consequently not increase errors because of weights with a suboptimal fit to unknown data.

The previous discussions indicated that especially combinations including higher numbers of forecasts are prone to small changes of the error characteristics. Thus, the shrinkage level required for robustness can be expected to increase with the number of forecasts. To provide insights into this aspect, Figure 4.13 shows the robust shrinkage factor for the M3 error covariance matrices, indicating to which extent OW must to be shrinked to achieve a certain degree of robustness. For the example, a training sample size of $n = 50$ is assumed and robustness parameters are set to $r = 0.05$ and $v = 0.1$. Consequently, robustness against an absolute change of an error correlation by 0.05 and a relative change of an error variance by 10 % is desired. The figure shows the empirical cumulative frequencies of robust shrinkage factors (the lowest shrinkage factor leading to lower expected error variance than SA under changes up to the desired ro-

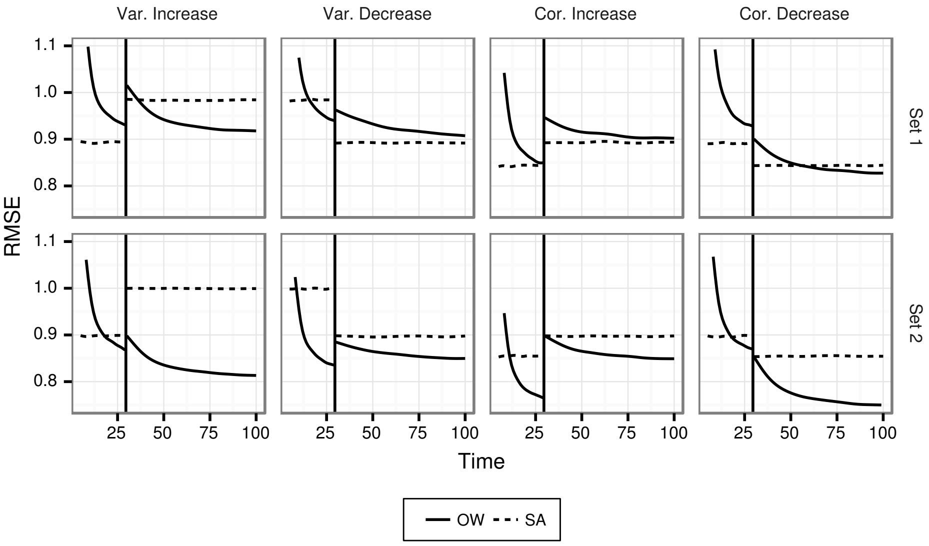


Figure 4.12: Average RMSE per time interval before and after the structural break (vertical line) for different parameter sets and treatments. While the same combination in many cases performs best before and after the structural break, the superior combination in some cases changes from OW to SA. The performance of OW improves again after the break since observations with the new error characteristic are included into the weight estimation.

bustness degree) for selected numbers of forecasts. The three plots in the figure present results for robustness only against changes of error correlation, of error variance, and of both characteristics.

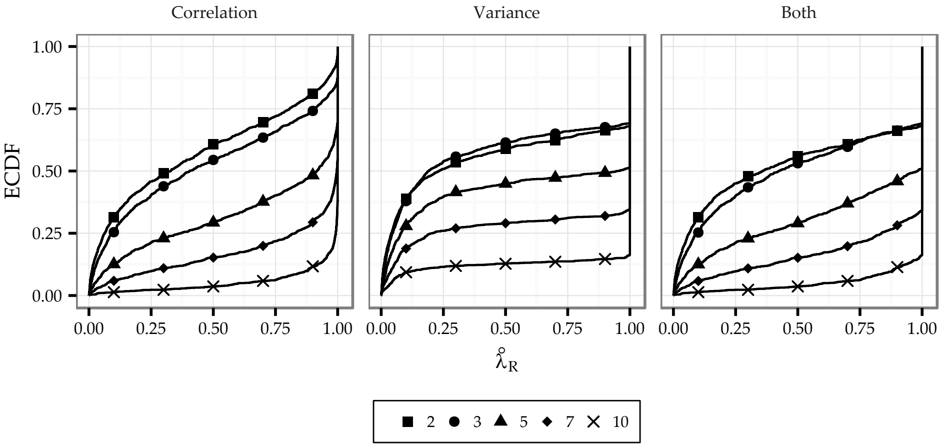


Figure 4.13: Empirical cumulative frequency of the robust shrinkage factor $\mathring{\lambda}_R$ required to ensure robustness against changes of correlation up to ±0.05 (left), of error variances changes up to ±10 % (center) and against both (right) for training sample size $n = 50$ and different numbers of forecasts k. Higher numbers of forecasts require using SA ($\mathring{\lambda}_R = 1$) in most cases to ensure robustness against both changes individually as well as against both.

The plots confirm the general tendency that the more forecasts are included, the higher the shrinkage required for robustness. The plots also show that the shapes of the cumulative frequencies are nonlinear: the higher the number of forecasts, the higher the frequency of robust shrinkage factors approaching one. While many cases allow a robust shrinkage factor below one for lower numbers of forecasts, an alternative to SA that is robust against variance or correlation changes, exists for less than 20 % of all matrices for ten forecasts.

As previously discussed, estimated weights tend to be more extreme for higher numbers of forecasts. Consequently, a stronger shrinkage must be used to achieve weights on a level comparable to the weights for fewer forecasts. To gain insights on the shrinked weights after applying the robust shrinkage factor, Figure 4.14 presents the distributions of the shrinked weights differing from equal weights (i.e. without the cases with $\mathring{\lambda}_R = 1$) for different numbers of forecasts. The forecasts are again ordered by difference of the weight estimate to equal weights in decreasing order. For instance $p = 1$ is the forecast with the strongest positive weight.

First of all, it can be noted that the number of cases where the weights after

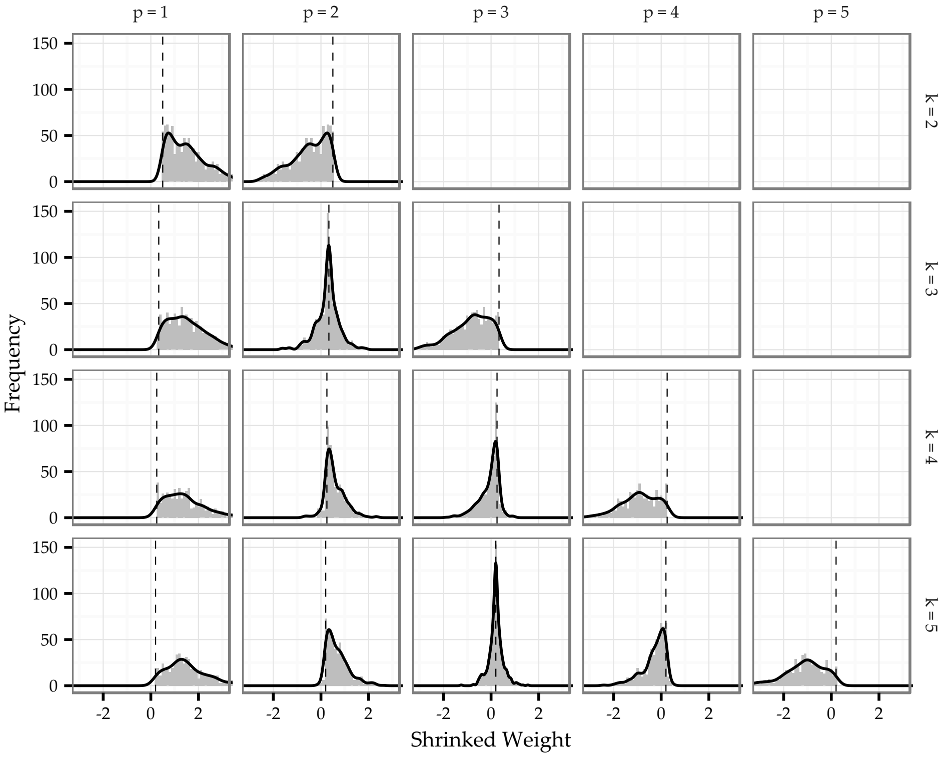


Figure 4.14: Histogram (gray) and density estimate (black line) of the weights shrinked with robust shrinkage, excluding cases with $\mathring{\lambda}_R = 1$, for different numbers of forecasts k. The distributions are displayed for individual forecasts and ordered by shrinked weight so that $p = 1$ has highest weight among the k forecasts. With increasing k, fewer robust combinations use weights other than equal weights. Furthermore, the shrinked weights are less differentiated.

robust shrinkage differ from equal weights decreases with increasing number of forecasts. This result is consistent with the previous discussions as combinations with more forecasts tend to be more prone to changes. Furthermore, the weights shrinked with robust shrinkage are less differentiated for higher numbers of forecasts. While weights with robust shrinkage are in a substantial number of cases higher than 2 or below 0 for $k = 2$, only few weights shrinked with robust shrinkage are as strong for $k = 4$ or $k = 5$. Consequently, lower numbers of forecasts allow a stronger fit to past errors while still ensuring robustness whereas higher numbers of forecasts require a weaker fit to past errors for robustness.

The earlier discussion and illustration of optimal shrinkage showed that the combined error variance with optimal shrinkage can be expected to continuously decrease with increasing number of forecasts. However, this is not necessary true when robustness is required as the discussion showed that strong shrinkage is required for robustness in case of many forecasts.

In order to analyze the impact of the (stronger) robust shrinkage factor on the performance of the combined forecast, Figure 4.15 presents the empirical distribution of the expected error variance with robust shrinkage (again with $r = 0.05$ and $v = 0.1$), relative to the $k = 2$ combination, for different numbers of forecasts. In contrast to the previous analyses, the forecasts are ordered by accuracy. The solid line again indicates the median and the light gray (dark gray) areas contain 75 % (50 %) of all observations per value of k. Since various different kinds of changes can occur, the worst case for robust shrinkage is considered, i.e., that error covariances do not change.

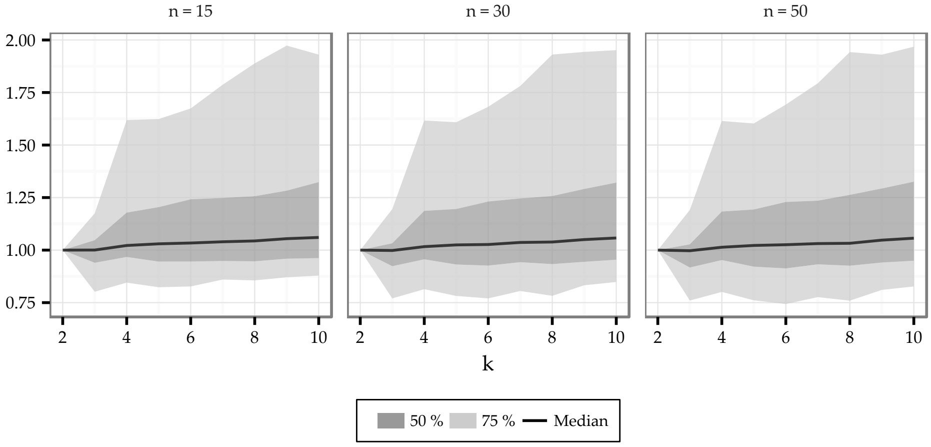


Figure 4.15: Distribution of the expected combined error variance with robust shrinkage when forecasts are ordered by accuracy, relative to the result for $k = 2$. The distribution is displayed for different training sample sizes. The median relative combined error variance is largely independent of the training sample size and increases with the number of forecasts, except for $k = 3$, which is slightly beneficial.

Comparing the distribution of the relative expected combined error variance for robust shrinkage to the results for optimal shrinkage (see Figure 4.7) reveals substantial differences. While optimal shrinkage results in a decreasing expected combined error variance, the combined error variance with robust shrinkage decreases only slightly for $k = 3$ and then increases. Consequently, using two or three forecasts is most beneficial in median when using robust shrinkage if the forecasts are ordered by accuracy.

Interestingly, the results differ substantially for a random ordering of the forecasts. Figures 4.16(a) presents the combined error variances, scaled by the corresponding results for $k = 2$. It can be noted that the distribution of the relative combined error variance in this case changes when increasing the number of forecasts. Left-tailed distributions can be observed when increasing k up to five, and then right-tailed shapes when including more forecasts in a combination. This can be seen as with $k \leq 5$ the 75th percentile is much closer to the median than the 25th percentile. This shape transformation can be observed for all training sample sizes considered. Hence, with lower number of forecasts, the mean relative combined error variance is lower than the median, while the mean exceeds the median for $k > 5$, significantly increasing with larger k. In contrast, the distributions are less beneficial when the forecasts are ordered by accuracy. While the distributions are largely symmetrical for small k, the distribution quickly changes to a right-tailed distribution with increasing k.

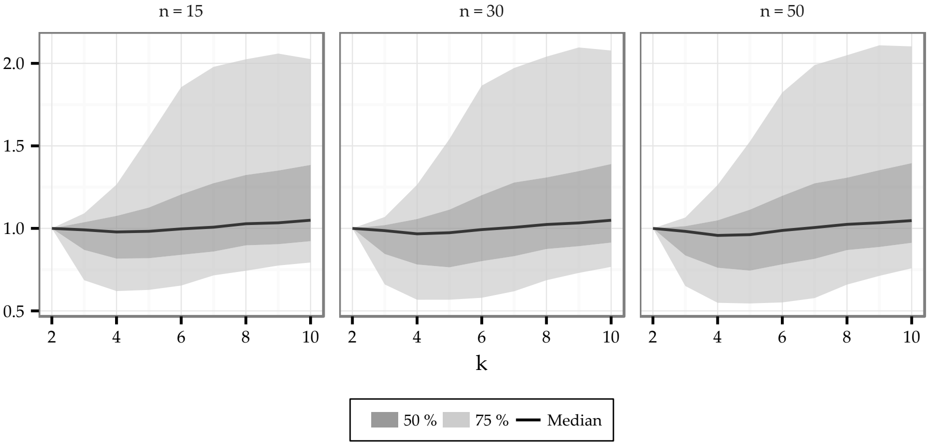
Overall, in contrast to the case with ordered forecasts, a range of three to five forecasts is beneficial when combining randomly ordered forecasts with robust shrinkage. However, as a result of the scaling, the previously discussed results of the two orderings cannot be compared directly. For this reason, Figure 4.16(b) additionally displays the combined error variances with random ordering scaled by the results for $k = 2$ in the ordered case. This scaling allows a direct comparison of the two orderings.

Interestingly, the plots indicate that the combined error variance with robust shrinkage is lower for a random ordering of the forecasts for the range of three to five forecasts. In other words, if three to five forecasts are combined, it is better to randomly select the included forecast instead of using the forecasts that can be expected to be most accurate.

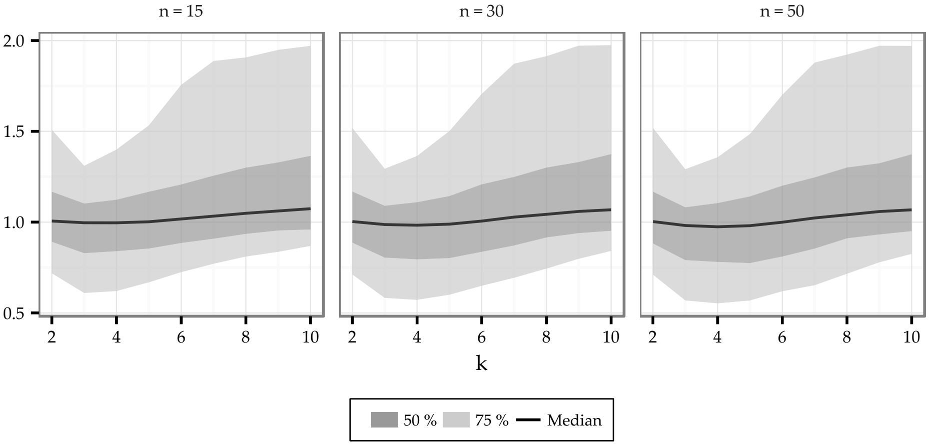
This finding might be counterintuitive at a first glance. However, it is in line with the literature on judgment aggregation (see Section 2.2), where diversity is often highlighted instead of accuracy. For instance, Davis-Stober et al. (2014) suggested selecting judgments that are as negatively correlated with each other as possible.

In order to provide insights into the diversity, Figure 4.17 compares the distribution of the mean correlation of the errors of one forecast with all other forecasts (which can be considered a measure of diversity) for the two different orderings. For instance the plot for $p = 2$ only considers the error correlation between the first two forecasts, either by random ordering or ordered by accuracy. For $p = 10$ the error correlations of the last with all other forecasts are considered.

The mean correlations are –besides random fluctuations– distributed very similarly for all forecasts in case of random ordering. In contrast, error correlations are highest for $p = 2$ for forecasts ordered by accuracy. The distributions are only similar for the two orderings for less accurate forecast (starting from $p = 7$). Thus, the ordering by accuracy also results in choosing forecasts with a higher error correlation when selecting a specific number of forecasts. However, as shown in the previous analyses and discussions, a higher error correlation results in weights



(a) Scaled by $k = 2$ for Random Ordering



(b) Scaled by $k = 2$ for Ordering by Accuracy

Figure 4.16: Distribution of the expected combined error variance with robust shrinkage in case of random ordering of the forecasts for different sample sizes n. The results in the upper plot are scaled by the combined error variances for $k = 2$ and random ordering whereas the lower plot uses scaling with the corresponding results for an ordering by accuracy. Both plots indicate that the combined error variance is lowest for a range of three to five forecasts and the advantage is most pronounced for $k = 4$. The lower plot furthermore shows that the combined error variance is lower for the random selection.

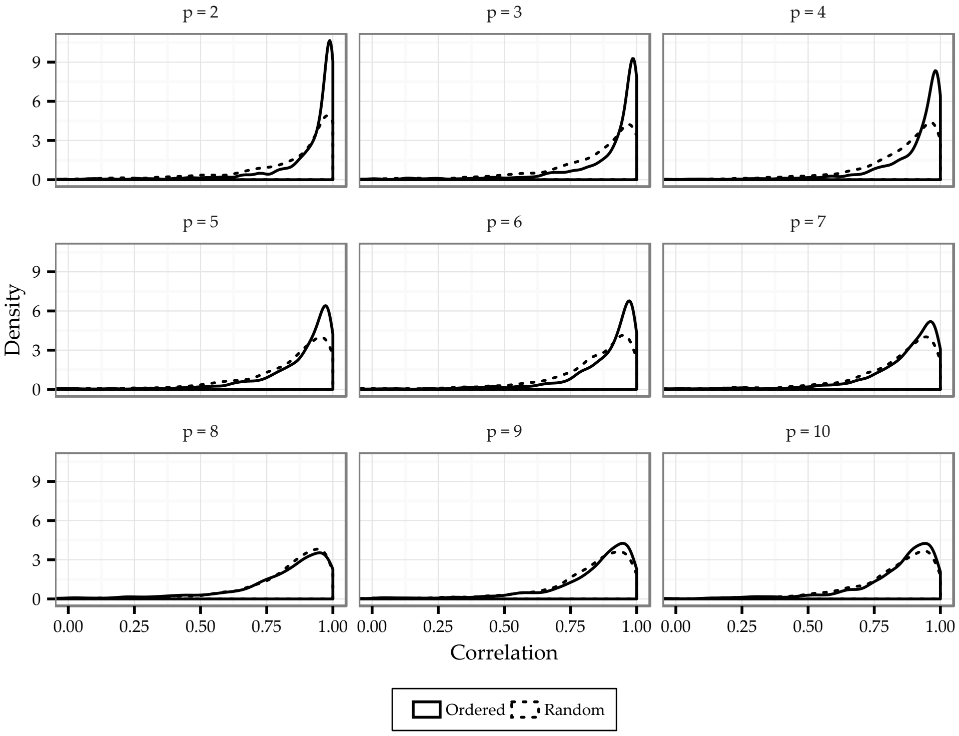


Figure 4.17: Distribution of the mean error correlation between forecast *p* and the previous forecasts, for random ordering and ordering by accuracy. The error correlations are strongest when the forecasts are selected by accuracy, as the distributions indicate higher correlations for this ordering up to approximately the seventh most accurate forecast.

that are more prone to structural changes and thus require a stronger shrinkage for robustness. This effect in turn increases the combined error variance.

In summary, the results indicate that using robust shrinkage instead of optimal shrinkage comes at substantial costs in terms of combined error variance. As higher numbers of forecasts require a stronger shrinkage for robustness, the expected combined error variance does not continuously decrease with increasing number of forecasts but has a minimum for a low number of forecasts. Thus, in contrast to the results for optimal shrinkage, only a limited set of forecasts should be included in the combination when aiming at robustness of the combination. Furthermore, randomly selecting forecasts for a combination with robust shrinkage can be expected to be more beneficial than selecting forecasts by accuracy. The higher diversity of randomly selected forecasts can be assumed as a key aspect leading to this result. The illustration and discussion of the analytical results thus provided valuable new insights into robustness in forecast combination and provides guidelines for practical application.

4.6 Conclusions and Limitations

In this chapter, a model of the expected out-of-sample error variance, consisting of a bias- and a variance-related component, of a forecast combination with optimal weights shrinked towards the simple average was introduced. The model was used to determine a shrinkage level that minimizes the combined out-of-sample error variance by balancing the bias and the variance component in order to adjust for the estimation uncertainty resulting from small training samples. Furthermore, critical changes of elements of the error covariance matrix were derived, which can for instance be used to determine how much elements are allowed to change for a combination to still outperform the simple average. The critical changes were finally used to introduce a novel robust shrinkage factor that ensures robustness against changes up to a definable extent.

The discussion of the optimal shrinkage level and the resulting expected combined error variance indicated that as much forecasts as possible should be included in the combination, as long as the available training sample is large enough. As little as 25 observations were found to be sufficient in most cases even when many forecasts, for instance 10, are combined. Especially for higher number of forecasts, only a moderate degree of shrinkage is required.

However, including as many forecasts as possible in a combination and not using strong shrinkage towards the average clearly contradicts the empirical results and recommendations in the literature. Consequently, estimation errors as a result of small training samples cannot completely explain the established result that simple averaging strategies perform as good as combinations with learned weights.

However, the illustration and discussion of the critical changes of error variances and correlations indicated that especially the combinations with larger numbers of forecasts are highly prone to small changes. The resulting expected increases in combined error variance can be large and easily result in a performance worse than the simple average.

The introduced robust shrinkage factor addresses this issue by increasing the shrinkage up to a level that ensures that the combination performs better than an alternative combination up to definable changes. The discussion showed that, as a result of the low initial robustness against changes, strong increases of the shrinkage level are required for more than three or four forecasts, often resulting in using the simple average. As a result, the combined error variance does not decrease as strongly with increasing number of forecasts. Only a range of three to five forecasts is beneficial over a combination of just two forecasts. Furthermore, a combination, which considers robustness by using the robust shrinkage level, is more beneficial when the forecasts included in the combination are selected randomly as a high diversity is ensured.

Thus, if structural changes have to be expected, a combination using the robust shrinkage factor introduced in this chapter should be considered. Furthermore, only a low number of forecasts should be included in the combination and the forecasts should not be selected by accuracy in order to ensure sufficient diversity of the forecasts.

The theoretical analyses as well as the illustration and discussion are subject to several limitations. First of all, the derived formulation of the expected out-of-sample combined error variance is based on the assumption of a multivariate normal distribution of the errors. Although this is a very widespread assumption, it might not be satisfied in practice and the errors might for instance follow a distribution with heavier tails, which would in turn increase estimation uncertainty and expected error variance. The distribution of the errors additionally assumed that the expected errors are zero, i.e., that the forecasts are unbiased. This assumption is especially likely to be violated if a judgmental forecast is included in the combination. However, although a bias of a forecast influences the combination and increases errors, the bias is reduced in a combination with unbiased forecast.

The derivation furthermore used the assumption that the weight estimates have no covariance with the errors of the forecasts included in the combination. As shown by Claeskens et al. (2016), a violation of this assumption can result in a bias of the combined forecast and an increased error, which is not considered in the analyses in this chapter, such as the optimal or robust shrinkage factor.

From a statistical point of view, the derived optimal shrinkage level is another random variable that has to be estimated. The uncertainty related to this estimation however increases the variance component of the error of the combined forecast. Thus using a more conservative than the calculated optimal shrinkage factor might be reasonable. This aspect is however not included in the analyses as it would introduce another layer of complexity.

The robust shrinkage as an alternative shrinkage level aims at increasing the robustness against structural changes. These changes of error characteristics can in practice occur in an arbitrary fashion for one or several of the characteristics and for one or several forecasts. However, the analyses of structural changes, and consequently the introduced robust shrinkage level, only considered one change at a time. Thus, the robust shrinkage level does not ensure robustness against multiple concurrent changes that might reinforce or negate each other.

Lastly, the derived models are theoretical ones and strongly rely on the error covariance matrix. Given a error covariance matrix, the analyses assume that weights are estimated from one sample, which is drawn from the distribution parameterized by the error covariance matrix, and then quantify the error variance resulting from applying these weight estimates. However, in practice, the error covariance matrix is unknown and only one sample of errors is available and used for parameter estimation. Thus, the model cannot be directly applied in specific settings. An application is only possible under the assumption that the estimate is equal to the population error covariance matrix. This assumption is clearly unlikely to be met in practice. However, the variance component included in the model, which is based on the sampling distribution, can be seen as a compensation of this aspect.

In order to analyze whether the analyses and models derived in this chapter not only provide guidelines for forecast combination, but can also be applied in practice despite potential violation of the discussed assumptions and limitations, an empirical evaluation is required. The corresponding empirical case study is presented in the next part of this work.

Part III

Application in Practice and Empirical Evaluation

Chapter 5

Case Study: Corporate Cash Flow Forecasting

THE previous chapters mainly focused on the theoretical properties of forecast correction and combination methods. In order to transfer the derived insights and the developed new models to practical applications, this chapter introduces a case study in corporate cash flow forecasting, where forecast correction and combination can be applied to judgmental forecasts.

Forecasts of future cash flows in different currencies play an important role in various management tasks in corporate finance. Cash flow forecasts are for instance used in liquidity management and foreign-exchange risk management. In liquidity management, the forecasts are used to anticipate cash shortages or surpluses in order to ensure solvency on the one hand and to limit cash reserves (which reduce profitability) on the other hand. In foreign-exchange risk management, cash flow forecasts are a basis for determining exposures resulting from business transactions in foreign currencies, which can then be hedged. Inaccurate forecast are an unreliable basis for financial plans. In liquidity and foreign-exchange risk management, inaccurate forecasts can lead to liquidity shortages and even insolvency, uncovered risks or increased hedging costs. Evidence for the importance of the accuracy of cash flow forecasts was for instance provided by Gormley and Meade (2007), who showed that the transactional costs for corporate-wide cash balance management using short-term cash flow forecasts strongly depends on the forecast accuracy.

In order to illustrate how judgmental cash flow forecasts are produced in practice, Section 5.1 introduces the sample company of the case study and empirical evaluation as well as the process in which the cash flow forecasts are produced. Subsequently, the forecast and realization data that is available from the sample company are introduced and described in Section 5.2. Lastly, the preprocessing applied to the available data is described in Section 5.3 before the data is used for an evaluation of forecast correction and combination methods in the next chapter.

5.1 Sample Company

In the sample company of the case study, cash flow forecasts of expected foreign currency denominated accounts receivables and accounts payables are used for instance for foreign-exchange risk management. The sample company is a multinational corporation with headquarters in Germany whereas operations are distributed worldwide. The company is, at the time of the case study, structured into three relatively independent business divisions with very different products and markets. Based on their business portfolios, the divisions are named "agricultural products" (AP), "health and pharmaceuticals" (HP) and "industrial materials" (IM) in this case study. Subsidiaries that cannot unambiguously be assigned to one division are grouped into the "diverse" (DIV) group. Over 300 separate legal entities with over 100,000 employees belong to the company. Thus, the company is strongly diversified and heterogeneous. Annual revenues amount to over 40 billion Euro, which mainly result from business in Europe, North America, as well as Asia and the Pacific Region.

As a result of the legal structure of the company, the financial management is centralized and the financial managers of the subsidiaries report to the central corporate financial controlling at the headquarters. The cash flow forecasts of the accounts receivables and accounts payables are consequently generated worldwide by the subsidiaries and are then submitted to the central corporate financial controlling. The accounts receivables result mainly from sales whereas the accounts payables are invoices from suppliers and other partners. Taken literally, accounts receivables and accounts payables are accruals rather than cash flows. As in most companies, historical data and forecasts of cash-ins and cash-outs are not available since the sample company's reporting systems are oriented towards revenues and expenses. The accounts receivables and accounts payables used in the analysis are however, for all practical purposes, comparable to cash-ins and cash-outs. For the sake of simplicity, the term cash flows is used even when referring to the accounts payables and receivables.

As the cash flow forecasts are produced by the financial managers of the subsidiaries based on available information, the forecasts are judgmental forecasts. However, as discussed in Section 2.1, judgmental forecasts are regularly found to be biased and strongly influenced by cognitive heuristics. The cash flow forecasts of the sample company are consequently likely to have substantial inaccuracies that can be reduced using statistical methods. It is furthermore likely that the biases substantially influence corporate planning and decision processes.

Despite the general awareness of the importance of accurate financial forecasts for corporate planning and control (Kim et al., 1998; Graham and Harvey, 2001), there is no research available that analyzes how judgmental corporate cash flow forecasts can be improved using statistical methods. This is in particular true for the correction of judgmental forecasts using systematic biases identified from past forecasts and the combination with model-based forecasts. The empirical application and evaluation consequently apply the forecast correction and combination approaches discussed in the previous chapters to the cash flow forecasts of the sample company. As a basis for the evaluation, the next section describes the available data as well as the preprocessing steps that were taken.

5.2 Available Data and Preprocessing

The data set available for the case study consists of forecasts and corresponding realizations and is provided by the sample company described in the previous section.

The forecasts are delivered by the subsidiaries at regular intervals and cover monthly intervals with differing forecast horizons of up to at least 12 months. As a schematic illustration, Table 5.1 shows the temporal structure of forecast deliveries for cash flows in 2012, with the months of forecast delivery (labeled F) and the month of the realization of the corresponding cash flows (labeled A). For instance, the forecasts for cash flows from January till March 2012 are delivered in March 2011 (with a horizon of four quarters), June 2011 (with a horizon of three quarters), and so on.

	2011												2012											
Horizon	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
4 Quarters			F										A	A	A									
						F										A	A	A						
									F										A	A	A			
											F											A	A	A
3 Quarters						F							A	A	A									
									F							A	A	A						
											F								A	A	A			
															F							A	A	A
2 Quarters									F				A	A	A									
											F					A	A	A						
															F				A	A	A			
																		F				A	A	A
1 Quarter											F		A	A	A									
															F	A	A	A						
																		F	A	A	A			
																					F	A	A	A

Table 5.1: Temporal structure of the deliveries of the judgmental cash flow forecasts in the case study in 2012. For different forecast horizons and months of actual cash flows (A), the delivery of the corresponding forecasts (F) is shown. For instance, the first forecasts for cash flows in January 2012 are delivered approximately four quarters ahead in March 2011 (1st row of the table) and are then revised in June 2011 (5th row of the table), in September 2011 (9th row of the table), and finally in November 2011 (13th row of the table), when the final forecasts for cash flows in January 2012 are delivered.

It should be noted that the structure of forecast deliveries has been changed several times during the time under study. For instance, the forecast deliveries in March and June were changed to February and May for practical reasons.

The forecasts and actual cash flow data are available for accounts payables and receivables. The data set spans actuals for the period from July 2008 to June 2014 and corresponding forecasts. The forecasts were delivered from November 2006 to May 2014 on an approximately quarterly basis.

The forecasts and actual values are available at a relatively fine grained level, which involves separate values for different partners. However, as business partners can regularly change, the data is aggregated to only differentiate between external and internal partners. As a result, the forecasts and actuals are available per subsidiary, currency, account payable or receivable, and external or internal partner.

Because of changes to the delivery structure, forecasts for some months are only available with horizons of up to 9 months. For reasons of consistency, forecasts with longer forecast horizons are excluded from the empirical evaluation. Furthermore, some of the available time series have only a short history and are thus not well suited for a thorough evaluation of the different forecast correction and combination approaches where parameters have to be estimated using past observations. At most 6 observations with zero or non-existing values are allowed within the first five years of the data and none within the last year as it used as evaluation period. Time series that do not meet these requirements are excluded from the evaluation.

Lastly, as illustrated in Table 5.1, forecast are produced approximately once every quarter in an irregular manner. For each actual value, only up to five unique forecasts exist. As a consequence, when considering the forecast horizon in months, relatively few observations exist per horizon. To circumvent this issue, the forecasts are considered up-to-date until a new forecast is available, i.e., forecast horizons that do not exist for a particular actual value are filled in with the last available forecast. This approach closely resembles practical applications, where a forecast is used until a revised one is available.

In order to provide more detailed insights into the data available after the preprocessing and its characteristics, the next section provides descriptives and furthermore classifies the data into categories as a basis for the analyses in the evaluation.

5.3 Data Characteristics and Classification

The data set used for the empirical evaluation results from applying the preprocessing steps described in the previous section to the available data set. The resulting time series data spans a period of six years and consists of 175 actuals time series and corresponding forecasts from 41 different subsidiaries and in 27 different currencies. Since several forecasts are available for each actual item, over 113,000 forecasts can overall be analyzed. The data statistics per division are presented in greater detail in Table 5.2.

	All Divisions	AP	HP	IM	DIV
Period	07/2008 – 06/2014				
Length of actual time series	72 months				
Subsidiaries	41	6	12	6	17
Currencies	27	10	21	15	5
# Actual Time Series	175	30	68	23	54
# Actuals	12,600	2,160	4,896	1,656	3,888
# Forecast Time Series	1,575	270	612	207	486
# Forecasts	113,400	19,440	44,064	14,904	34,992

Table 5.2: Descriptives of the available time series in the case study by business division. Overall, more than 12,000 actual values and corresponding forecasts with different horizons are available. Most observations are available for the HP and DIV business divisions.

The time series are likely to have different characteristics per business division because of their specific business characteristics. Companies in AP produce a broad spectrum of agricultural supplies and therefore largely depend on agricultural cycles, i.e., a yearly cycle of seeding, pest and insect protection, and harvesting. In contrast, IM develops and produces industrial materials. IM consequently depends on orders from manufacturing companies, which depend on the global economy and therefore on macroeconomic uncertainty. HP researches and produces health related products and pharmaceuticals, which do not (or only weakly) depend on the economy or annual cycles. These differences regarding dependencies of the business divisions' cash flows on seasonality and macroeconomic developments can also be seen in Figure 5.1, where the characteristics of the actuals time series and the judgmental forecasts are summarized.

The left plot in the figure illustrates the different variability of the time series in the different business divisions. One boxplot shows the coefficient of variation in one division, where one observation corresponds to the coefficient of one actual time series from an entity belonging to that division. The variability of the time series is on comparable levels except for the division AP, where the time series fluctuate more strongly.

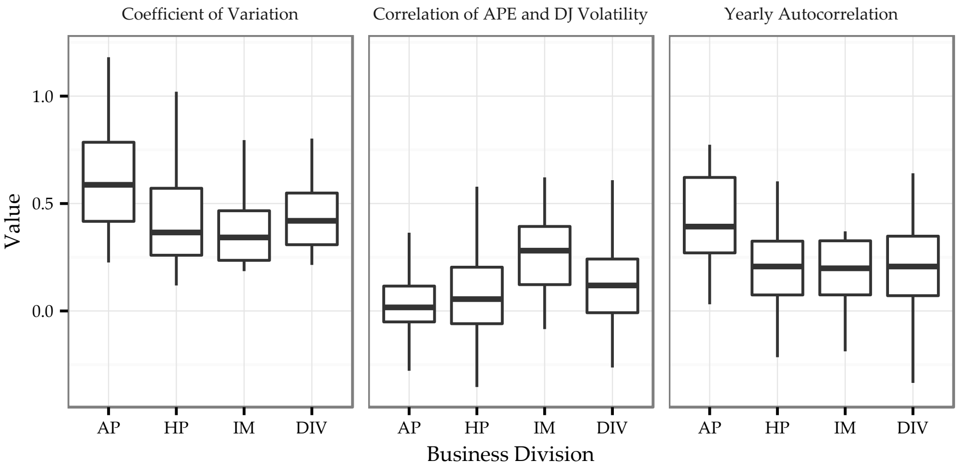


Figure 5.1: Characteristics of the actual and error time series by business division. The variability of the time series (measured by the coefficient of variation) is increased for AP, which is however largely explained by the yearly autocorrelation that is increased because of the dependence on annual agricultural cycles. The errors of forecasts for cash flows in business division IM show substantial correlation with the volatility of the Dow Jones Index, which is commonly used as a proxy of macroeconomic uncertainty.

The middle plot in the figure shows the correlation of the forecast errors (measured by the absolute percentage errors, APE) and the volatility of the Dow Jones (DJ) Index, which is often used as a proxy for macroeconomic uncertainty. A strong relationship is only found for cash flows in division IM but not in other divisions. This is a result of the division's dependency on the global economy.

The right plot in the figure shows the distribution of the autocorrelation of actual cash flow time series with a lag of twelve months (yearly seasonality) by business division. AP exhibits the largest autocorrelation in its cash flows, as can be expected from the agricultural business. The high autocorrelation also explains the high volatility quantified by the coefficient of variation. The other two divisions and DIV overall show little autocorrelation.

The different business characteristics highlight the high diversity of the time series included in the empirical evaluation. Even though the data only stems from one corporation, the results can be expected to be largely generalizable because of the heterogeneity of the time series.

As time series characteristics are still relatively heterogeneous within the business divisions, a more detailed classification of the time series is of interest for the analyses. Some approaches, especially in forecast correction, can be expected to perform better for specific types of the time series, such as time series with seasonality.

	Time Series Classification			
Division	None	Trended	Seasonal	Both
AP	6	6	10	8
HP	16	33	4	15
IM	6	10	0	7
DIV	14	24	1	15
Sum	42	73	15	45

Table 5.3: Number of time series per type by division as a result of the time series classification. For AP, most time series are either seasonal or seasonal and trended, whereas trended time series are most frequent for the other divisions.

In order to evaluate this dependency, the available time series are classified into the four following types. (i) *Trended* if a KPSS unit root test (Kwiatkowski et al., 1992) indicates non-stationarity. (ii) *Seasonal* if the yearly autocorrelation is strong (above the 66 % quantile, 0.324 for the available time series). (iii) *Trended & Seasonal* if a time series is trended and seasonal according to the criteria above. (iv) *Stationary* if a time series does not match any of the previous types.

The resulting classification is shown in Table 5.3, where the number of time series of the different types is displayed by business division. Overall, the classification results in 73 trended (42 %), 15 seasonal (9 %), 45 trended and seasonal (26 %) and 42 stationary time series (24 %).

As some approaches might be more robust against time series with higher variability, an additional label is generated indicating whether a time series of actual values has a high variability. For this purpose, the 66 % quantiles of the coefficient of variation per time series type are used. In the case study, the threshold is approximately 0.5 for all types of time series except for seasonal time series, where the threshold is 0.76. Besides measuring the variability of the time series, the metric can furthermore indicate whether an actual time series is likely to contain an outlier, as the coefficient of variation is often strongly increased for time series with outliers.

Overall, a large data set of judgmental forecasts and corresponding realizations from a real-world application is available for the empirical evaluation. The data set is very heterogeneous regarding characteristics of the time series as well as regarding biases and error patterns in forecasts as they were produced by a variety of experts. Thus, the data set is a solid basis for the empirical evaluation of forecast correction and combination approaches, which is presented in the next chapter.

Chapter 6

Empirical Evaluation

THE case study in corporate cash flow forecasting introduced in the last chapter illustrates the importance of the accuracy of judgmental forecasts. As a considerable history of past forecasts and realizations are available, forecast correction as well as combination can be applied in order to improve the accuracy of the judgmental forecasts. In this chapter, the available data is used to evaluate the different approaches introduced in this work. First, Section 6.1 introduces the research design, which is used in the empirical evaluation. Subsequently, the results on forecast correction and combination are discussed separately in Sections 6.2 and 6.3. The two approaches are then compared in Section 6.4. Lastly, conclusions regarding the available mechanisms as well as limitations of the empirical evaluation are discussed in Section 6.5. Furthermore, potential starting points for future improvements in the integration of human judgment and statistics in forecasting tasks are derived.

6.1 Research Design

For a thorough evaluation of the forecast correction and combination methods, the design of the empirical evaluation should resemble real-world applications as closely as possible. Thus, the various parameters required for the different models have to be estimated from a training data set and then be applied to an independent evaluation data set, which is not considered in the estimation. The evaluation data set can then be used to measure and compare the performance of different methods.

Of the six years of available data in the case study, actuals (and corresponding forecasts) of the last 12 months (07/2013 – 06/2014) are used as evaluation data, where the performance of different forecast correction and combination approaches are compared. The evaluation uses a rolling approach where models are learned for the first data point of the evaluation data set using the complete available history of available data prior to the evaluation data point (while considering restrictions resulting from the forecast horizon) as training data. The data point is then added to the training data and the procedure is repeated until the complete evaluation data set is processed.

It should be noted how forecast horizons influence the available training data. If for instance models for July 2013 are learned, all data until June 2013 is available as training data for a forecast horizon of one month whereas only data until April 2013 is available in case of a forecast horizon of three months. This restriction ensures comparability to applications in practice, where actual values are only available after the realization date has passed.

This evaluation setting is used for the evaluation of forecast combination as well as forecast correction models. The different treatments regarding models and their parameterizations, which are compared in the evaluation, are introduced next. Furthermore, a short introduction and overview is given how the model-based forecasts required for a forecast combination are calculated.

6.1.1 Treatments Regarding Forecast Correction

In the empirical evaluation, different configurations of the proposed extended forecast correction method, as introduced in Section 3.4, can be applied. A first parameter is the estimation method used in the linear regression. OLS as well as LAD are included in the evaluation, where the former can be expected to result in a better fit to the training data whereas the latter ensures a higher robustness against outliers.

In the estimation procedure, different weights of past observations can be used. The standard approach is to assign equal weights to all past observations. As a simple alternative, exponentially decreasing weights with a fixed discount factor can be used. Reasonable choices of discount factors clearly depend on specific time series and no universally valid recommendation exists. Goodwin (1997) used discount factors ranging from 1.0 to 1.2 with increments of 0.01. In the evaluation, a wider range of weights between 1.0 and 2.0 with increments of 0.01 is considered. A more complex alternative is learning weights by minimizing the error in the pseudo out-of-sample evaluation described in Section 3.1, as proposed by Goodwin (1997) and .

In order to address non-stationarity of time series, two methods that ensure stationary of trended time series are included. One the one hand, simple differentiation of time series with a lag equal to the forecast horizon is used. On the other hand, a trend is estimated and removed using an OLS linear regression. The common approach of using log-returns of time series is not included in the

Treatment:	Estimation Method	Weighting	Transformation	Breakpoint
Scope:	WLS LAD	Equal Weigthing Fixed Weighting (1.0 - 2.0, step 0.01) Weight Learning	None Differentation Detrending STL	No Breakpoint Breakpoint
Count:	2	3	4	2

Table 6.1: Treatments regarding different forecast correction methods included in the evaluation. Two different estimation methods are used with either equal weighting, fixed exponential weighting or weights learned from past observations. A breakpoint detection is used for some models. Before the forecast correction methods are applied, different transformations to the time series are used to ensure stationarity.

empirical evaluation since applying this approach requires time series having no zero values, which is not the case in the available data. A seasonal component can be addressed by using STL (Cleveland et al., 1990) to identify (and remove) systematic trend and seasonal components, which can then be extrapolated to the evaluation sample with exponential smoothing.

Lastly, models can be differentiated into ones that detect and incorporate potential breakpoints, as proposed in Section 3.4, and ones that ignore structural changes.

Table 6.1 summarizes the different treatment variables and their values. Overall, 48 treatments regarding the different model configurations are available when ignoring the different fixed weights.

Experiments are conducted using *R* (R Core Team, 2015) and the provided methods for (weighted) linear regression and STL. For LAD, the *quantreg* package (Koenker, 2015) is used. Breakpoints are detected using the corresponding function of the *strucchange* package (Zeileis et al., 2002).

It should be noted that a dedicated model is learned for each time series and forecast horizon individually and not for groups of time series. Time series characteristics and error patterns can differ substantially between time series, thus learning correction models for multiple time series is not reasonable. Likewise, error patterns can differ between forecast horizons (most obviously, error variances decrease with decreasing forecast horizon). Consequently, a differentiation between forecast horizons is required.

In contrast to forecast correction, forecasts besides the judgmental one are required for a combination of forecasts. Thus, before the treatments regarding forecast combination are introduced, the calculation of the model-based forecasts is discussed next.

6.1.2 Calculation of Model-Based Forecasts

While the previously described evaluation setting can be used for evaluating forecast correction approaches, which only use past forecasts and corresponding errors for model estimation, an evaluation of forecast combination approaches requires a more complex setting. A combination of forecasts obviously requires at least two available forecasts. Whereas judgmental forecasts are already available per time series, model-based forecasts have to be calculated additionally. Since most forecast combination approaches use past error variances and correlations of forecasts to estimate weights, model-based forecasts not only have to be produced for the evaluation data but also for a large part of the training data.

For the evaluation, model-based forecasts are produced starting from 07/2011 where, depending on the forecast horizon, up to three years of training data are available for estimation of the forecasting model. Consequently, between 14 and 35 past errors are available per forecast for estimating combination weights, depending on the forecast horizon and the position of the forecast in the evaluation sample.

As statistical time series forecasting models, autoregressive integrated moving average (ARIMA) and damped trend exponential smoothing (DTES) models are used. Both models are often recommended in the literature, as discussed in Section 2.2. Both models are shortly introduced below in order to illustrate the basic idea of the time series forecasting models.

An ARIMA model can first be decomposed into an autoregressive (AR) and a moving average (MA) model. In an AR model, generally denoted by $AR(p)$, each observation A_t of a time series is modeled as a weighted (linear) sum of the last p values, as shown in 6.1. The reasoning behind this modeling is that the last p values, which are weighted with ϕ, are expected to contain most information about the next value. An additional constant c is used and only a random unsystematic component ϵ_t remains.

$$A_t = c + \phi_1 A_{t-1} + \cdots + \phi_p A_{t-p} + \epsilon_t \tag{6.1}$$

In contrast, a MA model, denoted by $MA(q)$, uses the last q errors (instead of actual values) for modeling A_t. For this purpose, the last q errors are weighted with a vector θ and an additional shift μ is introduced, as shown in Equation 6.2. It should be noted that the error terms $\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}$ are not observable. The interpretation of a MA model is that, beyond μ, the value of A_t only depends on the last errors, i.e., white noise disturbances. This modeling is chosen to allow the disturbances to result in deviations from the mean, which then fade over time.

$$A_t = \mu + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{6.2}$$

The introduced AR and MA models can be combined in an $ARMA(p, q)$ model, as shown in 6.3, which considers autoregressive dependencies as well as disturbances away from the mean influencing the time series.

$$A_t = c + \phi_1 A_{t-1} + \cdots + \phi_p A_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{6.3}$$

The previous definitions assumed stationarity of time series, i.e., that characteristics such as the mean and the variance to not change over time. If a time series is not stationary, deriving a stationary time series with values A'_t is required by differencing the observations d times, as shown in Equation 6.4. The ARMA model, which uses observations differenced d times instead of the original observations, is then denoted $ARIMA(p, d, q)$.

$$\begin{aligned} d = 0: &\quad A'_t = A_t \\ d = 1: &\quad A'_t = A_t - A_{t-1} \\ d = 2: &\quad A'_t = (A_t - A_{t-1}) - (A_{t-1} - A_{t-2}) \\ \ldots &\quad \ldots \end{aligned} \tag{6.4}$$

All of the different models depend on parameters such as ϕ and θ, which have to be estimated from past observations. After the parameters have been estimated, the model can be used to forecast, i.e., extrapolate, future developments of the time series by using the observations up to A_t to predict A_{t+1}.

However, before the parameters can be estimated, the model (i.e. the values of p, d, q) has to be selected. This step can either be done by a forecasting expert focusing on time series modeling or by automatically estimating various models and selecting the one with the best performance, considering model complexity, on the basis of a comparison using the Akaike Information Criterion (AIC), as proposed by Hyndman and Khandakar (2008).

Although it was developed from exponential smoothing models, damped trend exponential smoothing corresponds to a specific ARIMA model, $ARIMA(1, 1, 2)$, which is shown in Equation 6.5. The idea of damped trend exponential smoothing is to extrapolate the local trend $A_{t-1} - A_{t-2}$, but to dampen the trend using the last disturbance terms. This "introduces a note of conservatism", as stated by Gardner and McKenzie (2011).

$$A_t = A_{t-1} + \phi_1 \left(A_{t-1} - A_{t-2}\right) - \theta_{t-1}\epsilon_{t-1} - \theta_{t-2}\epsilon_{t-2} \tag{6.5}$$

In the case study, DTES and ARIMA forecasts are calculated for all time series. Thus, two model-based forecasts are available in addition to the judgmental forecast, which allows different combinations of the forecasts. The combinations as well as the different combination methods in the empirical evaluation of forecast combination are introduced next.

6.1.3 Treatments for Forecast Combination

A forecast combination, as analyzed in Chapter 4, can differ by the forecast included in the combination and by the approach used to determine the weights of the forecasts. In the evaluation, a judgmental forecast produced by an expert (denoted Exp) is already available. Alternative model-based forecasts are produced using the previously introduced DTES and ARIMA models. Overall, three forecasts are available, allowing three different combinations with two forecasts and one combination with three forecasts.

As standard weighting approaches of the forecasts, the simple average (SA) and Bates and Granger's optimal weights (OW) are considered. Beyond these two in terms of the bias–variance trade-off extreme variants, weights with optimal and robust shrinkage, as introduced in Chapter 4, are used. Robust shrinkage requires additional parameterization, error variance or correlation changes against which the combination should be robust. For relative changes of the error variances, values indicating low (10 %), medium (30 % and 60 %) as well as high robustness (90 %) are included in the evaluation. Absolute changes are used for the error correlation, with different values for low (0.1), medium (0.3 and 0.6) as well as high (0.9) robustness.

Table 6.2 summarizes the different treatment variables and their values. Overall, 16 basic treatment combinations are included in the evaluation. When considering the different variants of robust shrinkage, the number of combinations increases to 76.

Experiments are conducted using *R* (R Core Team, 2015) and the model-based forecasts are produced using the auto-ARIMA and Damped Trend Exponential Smoothing models in the *forecast* package (Hyndman, 2015).

It should again be noted that a dedicated combination model is learned for each time series and forecast horizon individually and not for groups of time series or multiple horizons as time series characteristics and error patterns can differ substantially between time series.

		Weighting Scheme		
Treatment:	Forecasts	Approach	Robustness Variance	Robustness Correlation
Scope:	Exp, DTES	SA	none	none
	Exp, ARIMA	OW	none	none
	DTES, ARIMA	Opt. Shrinkage	none	none
	All	Rob. Shrinkage	0.1, 0.3, 0.6, 0.9	0.1, 0.3, 0.6, 0.9
Count:	4		19	

Table 6.2: Treatments regarding different forecast combinations. Three different forecasts are available, resulting in four different combinations with two or three forecasts. The available forecasts can be combined with SA, OW or with OW shrinked towards SA using the two introduced shrinkage levels.

Based on the introduced experimental setting of the empirical evaluation, the results on forecast correction and combination using different approaches are presented in the next section.

6.2 Results on Forecast Correction

The evaluation setting and the heterogeneous time series data introduced in the previous sections allows a thorough evaluation of forecast correction. The results of the evaluation of the different correction methods and parameterizations are presented and discussed in this section.

As a metric of the performance of the different forecast correction approaches, the MSE is used as it is the metric optimized by the correction methods. However, MSE values are likely to differ by orders of magnitude between time series as a result of the different scale of the time series and, consequently, the errors. Thus, to ensure comparability, the relative difference between the MSE of the corrected forecast and of the original judgmental forecast is used. This metric furthermore allows a direct assessment of the improvements over the original forecast. A negative relative MSE difference indicates that the corrected forecast performs better than the original forecast whereas a positive value indicates that the accuracy is negatively influenced.

As the data transformation method used to ensure stationarity of the time series as well as the time series characteristics can be expected to influence the results most fundamentally, both aspects are included in all analyses.

Since the various other different treatments regarding forecast correction methods allow a great variety of parameterizations, the results are analyzed in multiple steps. In a first step, different estimation methods, i.e., penalty functions, are compared. Then, the weighting schemes under study are compared for different data transformations and time series types. Subsequently, methods that consider breakpoints are compared to methods without explicit detection and incorporation of potential breakpoints. Lastly, all results are compared in a regression analysis in order to identify potential multivariate effects as well as to examine the significance of the results.

6.2.1 Estimation Method

Forecast correction methods can be expected to be prone to outliers in the data since a linear regression is used. Alternative estimators promise to increase robustness against potential outliers by reducing the weight of extreme deviations.

In order to evaluate the influence of the chosen estimator on the result of the forecast correction, Figure 6.1 compares results for OLS and LAD estimation. For this purpose, the relative MSE differences are presented for different data transformations and by type of time series. Furthermore, results are presented separately for time series with a high variability (measured by the coefficient of variation, CoV) as these are more likely to contain outliers. For this first analysis, approaches with breakpoints are not considered whereas equal as well as learned weights are included in the analysis.

Whether improvements over the original forecast can be achieved strongly depends on the time series characteristics in combination with the data transformation used as well as on the variability of the time series.

For time series with low or medium variability, the transformation methods overall match the time series well. In case of stationarity, no transformation is required, detrending or differentiation is beneficial for trended time series. For seasonal or trended and seasonal time series, a transformation using STL is required. One interesting result is that for seasonal time series, no transformation also results in substantial improvements. This indicates that the original forecast correction method by Theil (1966) is better suited for seasonal non-stationarity than for regular non-stationarity.

The differences between outcomes when using OLS or LAD are relatively small, both estimation methods result in approximately equal performance. In most cases, differences are however slightly in favor of OLS. Only in few cases for time series with high variability, an advantage of using LAD can be noted. For these time series, LAD is always of advantage regarding the median when using STL or when using detrending for seasonal or trended and seasonal time series.

In comparison to the original judgmental forecast, the median MSE difference is negative (indicating improvements) for almost all types of time series and transformations for time series with low or medium variability. In contrast, the

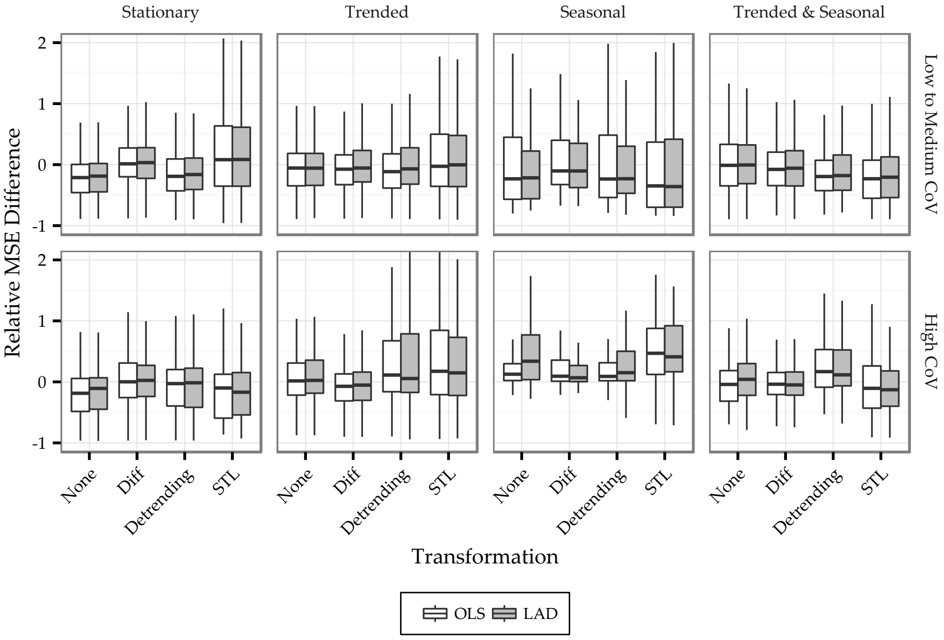


Figure 6.1: Distribution of the MSE difference to the original forecast for correction using OLS and LAD estimation. Results are presented by time series type and data transformation as well as for different degrees of variability of time series. LAD only in few cases performs better than OLS. The plot furthermore indicates clear guidelines on which data transformation to use for which type of time series.

forecast accuracy is negatively influenced (positive MSE difference) in many cases with high variability, especially for seasonal time series.

Overall, using OLS estimation can be recommended while using no transformation for stationary time series, differentiation for trended time series, and STL for trended and seasonal time series. In case of seasonal time series, STL should be used, except for cases with high variability, where a correction can be expected not to be beneficial.

The analyses on which these guidelines are based however included results for equal weighting as well as learned weights and did not consider breakpoints. These aspects are discussed next. All of the following analyses are limited to results for OLS estimation as LAD estimation only showed performance superior to OLS in few cases.

6.2.2 Weighting of Past Observations

In contrast to LAD, which aims at increasing robustness against outliers, a weighting of past observations aims at increasing robustness against changing

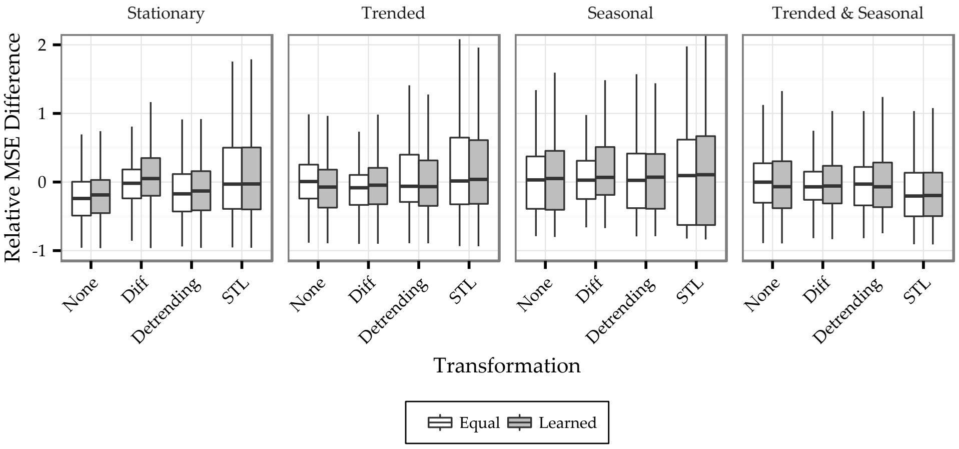


Figure 6.2: Distribution of the relative MSE difference to the original forecast for correction using equal and learned weights. Results are presented by time series type and data transformation. Learning weights is slightly beneficial when no preprocessing is used and time series at least have a trend. No clear advantage can be noticed in most other cases.

biases in forecasts. Besides an equal weighting of past observations, an exponential weighting can be learned in a pseudo out-of-sample evaluation on available training data. Detected breakpoints are, in a first step, excluded for the evaluation of the weighting schemes.

Figure 6.2 compares the outcomes for the two weighting approaches, differentiated by type of time series and data transformation. Surprisingly, no clear general performance increase of the learned weights can be noticed. Learning weights is only of clear benefit when using no preprocessing in case of trended or trended and seasonal time series. In most other cases, learning weights increases errors or results in both approaches performing approximately equal.

The observation that learning weights is beneficial for no preprocessing and time series with at least a trend likely results from issues with the standard correction method in case of non-stationary time series. If equal weighting is applied, the estimated mean is far away from the last actual value of the time series and consequently also differs substantially from the actual corresponding to the forecast that is corrected. By learning weights, strong exponential weights can be learned, which result in the estimated mean being closer to the last actual value and thus to future values. The more accurate estimate of the mean is clearly a better basis for the correction.

Overall, learning weights has no clear benefit over using equal weights if non-stationarity is treated appropriately. In order to assess this result in greater detail, Figure 6.3 provides an additional analysis for the treatment with exponential weighting with fixed discount factor and, consequently, without breakpoints. The relationship between the relative MSE difference and the exponential weighting factor is displayed for the different time series types and data transformation methods. The solid line indicates the median relative MSE difference whereas the dark gray (light gray) areas contain 50 % (75 %) of all observations per exponential weighting factor γ.

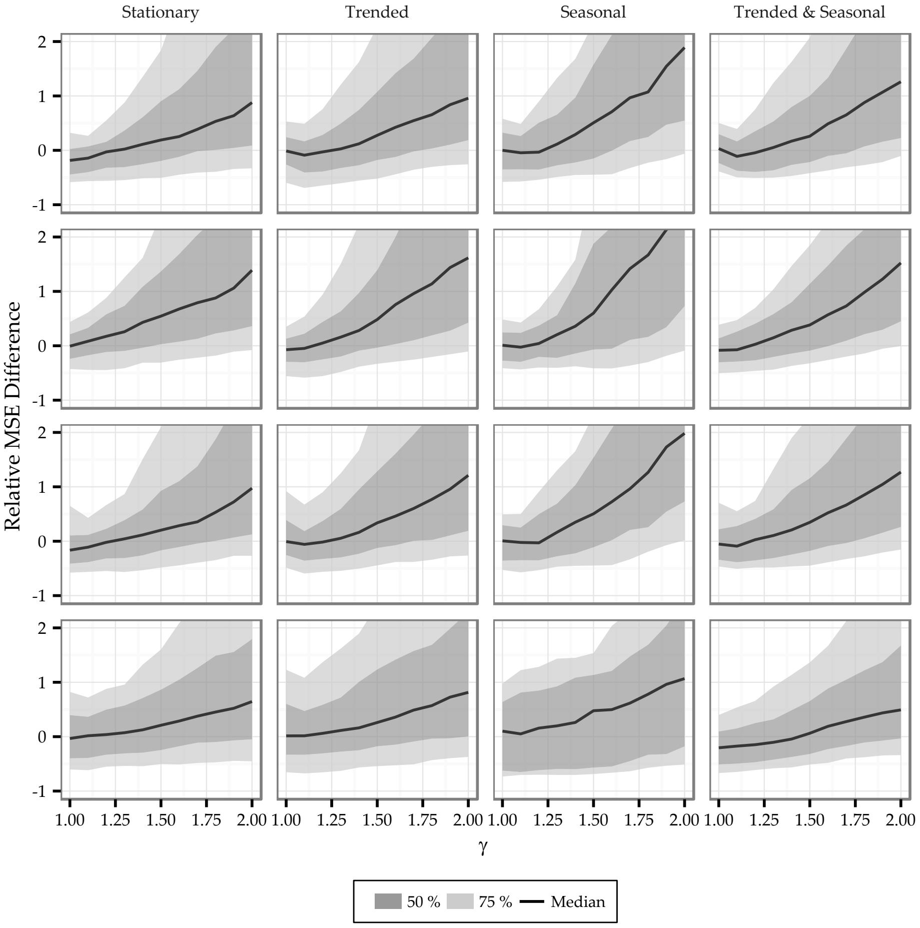


Figure 6.3: Distribution of the relative MSE difference to the original forecast for correction using exponential weighting with fixed discount factor γ. Results are presented by time series type and data transformation. An exponential weighting with discount $\gamma \approx 1.1$ is of clear benefit when no preprocessing is used. In other cases, equal weights perform as least as good as exponential weights if non-stationarity is addressed appropriately.

If no data transformation is applied, discount factors are beneficial (in comparison to the original forecast) up to 1.25, when taking the median of the relative

MSE difference as a criterion. In these cases, except for stationary time series, the MSE difference a U-shape, indicating that there is in fact a discount factor that outperforms equal weights. The most promising exponential weighting is with a moderate discount of approximately 1.1, for which the weights decrease relatively slowly but virtually no weight is assigned to observations older than 1-2 years.

However, if a data transformation is used, the plots only exhibit U-shapes few cases, indicating that an equal weighting is in most cases most beneficial across time series, especially if the non-stationarity of time series has been treated appropriately.

Overall, there is only a global optimum of the discount factor unequal to no discount if non-stationarity is ignored. Thus, learning weights is only of clear benefit when no transformation is used. There might however be beneficial discounts for individual time series that could improve accuracy in these cases. However, the previous results, which showed that learning weights often decreases accuracy, indicate that advantageous discounts are not trivial to identify and do not improve the bias–variance trade-off as the increase of the variance component is only in few cases balanced by a reduction of the bias component.

Thus, learning weights can in general not be recommended if –in contrast to existing evaluations in the literature– a reasonable transformation of the data is applied that ensures stationarity. However, the extended model proposed in this work additionally introduced new weighting approaches if breakpoints are explicitly detected and incorporated. The results for these approaches are analyzed next.

6.2.3 Approaches to Treating Structural Change

Structural changes or breaks can be implicitly addressed in the approach with learned weights by reducing the weight of old observations in order to minimize the weight of outdated observations. However, this also results in a very unequal distribution of the weights of more recent observations, which in turn increases the instability of the estimates. For this reason, the extended model included the possibility to explicitly incorporate potential structural breaks that are detected using established statistical means.

As a consequence, after the correction approaches that identify and treat potential breakpoints have been ignored in the previous analyses, Figure 6.4 presents results comparing methods with and without breakpoint. The results are differentiated by the general weighting approach as the proposed extended correction model considers identified breakpoints differently for equal and learned weights,

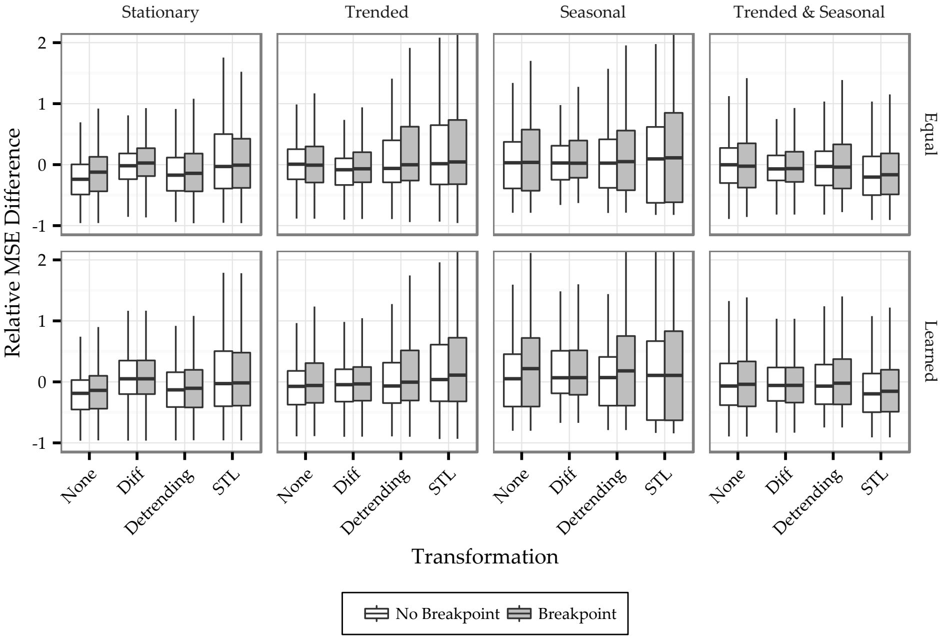


Figure 6.4: Distribution of the relative MSE difference to the original forecast for correction using methods with and without explicit identification and treatment of breakpoints. Results are presented by time series type and data transformation. Detecting and incorporating potential breakpoints is of no clear benefit in terms of median relative MSE difference and, in most cases, upper and lower quantiles.

as described in Chapter 3. The presented results are again restricted to OLS estimation.

Using approaches with breakpoints clearly does not improve median relative MSE difference, independently of time series type and data transformation method. Results differ only slightly for equal and learned weights, where identified breakpoints are considered in different ways.

Overall, either there are few time series with structural changes in the data set or the methods with breakpoints are not beneficial in terms of the bias–variance trade-off. Since the time series and forecasts of the case study are real-world data, it is very likely that bias changed over time or that one or several forecasters responsible for specific forecasts changed. Thus, it is unlikely that there are few structural changes within the available data.

In contrast, detecting breakpoints is clearly a complex task with substantial uncertainty. Detecting a false breakpoint can result in decreased training sample sizes and consequently increased instability of parameter estimates. This reasoning is confirmed by various additional analyses not shown in this work for reasons of comprehensibility. For instance using equal weights while reducing the weights of observations before an identified breakpoint using the parameter α did not result in increased accuracy. Thus identifying breakpoints that can be treated in a way that is beneficial for the correction is already a complex task. Furthermore, as shown in Chapter 3, the robustness of the correction methods is in many cases relatively high even without exponential weighting or explicitly considering potential breakpoints.

In summary, the analyses of the correction results clearly indicate that treating non-stationarity appropriately is of high importance. However, if non-stationarity is treated, learning weights is not required or beneficial. Likewise, explicitly detecting and incorporating breakpoints cannot be recommended as it seems to introduce too much uncertainty. In order to summarize the previous analyses into a comprehensive one and to assess the significance of the differences, a regression analysis is presented next.

6.2.4 Regression Analysis

The performance of forecast correction methods depends on a variety of aspects, as the previous analyses have shown. These include choices regarding the forecast correction model as well as the characteristics of the time series.

Both aspects are considered in the regression analyses of the relative MSE differences shown in Table 6.3. The results of four different regressions for the time series types are shown in the table. The basic correction model (no transformation of the time series, OLS estimation, equal weights, no breakpoints) for low to medium variability is used as baseline in the regression.

The intercepts in the regressions show that the base model already results in significant improvements for stationary and seasonal time series and low to medium variability. The improvements are strongest for stationary time series, where the standard forecast correction method without transformation of the time series could be expected to perform best, based on the theoretical analyses in Chapter 3. No accuracy differences to the original judgmental forecast are found for trended as well as trended and seasonal time series.

For stationary time series, applying any transformation to the time series significantly decreases the performance of the forecast correction. In contrast, for trended time series, using differentiation is significantly beneficial and essential for improvements in comparison to the original forecast, whereas all other transformations significantly reduce performance. Interestingly, no significant differences between transformations are found for seasonal time series. Lastly, for time series with trend and seasonality, using STL is most beneficial while using differentiation is also of significant advantage.

	Stationary	Trended	Seasonal	Trended & Seasonal
(Intercept)	-0.114^{***}	-0.004	-0.088^{**}	0.004
Detrending/No Transf.	0.071^{***}	0.091^{***}	0.004	0.026
Diff/No Transf.	0.156^{***}	-0.092^{***}	-0.040	-0.091^{***}
STL/No Transf.	0.216^{***}	0.141^{***}	0.047	-0.155^{***}
Learned/Equal W.	-0.001	-0.058^{**}	0.117^{**}	0.010
Detrending/No Transf. × Learned/Equal W.	0.011	0.028	-0.033	0.016
Diff/No Transf. × Learned/Equal W.	0.067^{*}	0.124^{***}	0.023	0.026
STL/No Transf. × Learned/Equal W.	0.010	0.068^{**}	-0.093	0.014
LAD/OLS	0.006	0.018	-0.073^{**}	0.023
High/Low to Medium CoV	-0.041^{*}	0.154^{***}	0.462^{***}	0.156^{***}
LAD/OLS × High/Low to Medium CoV	-0.006	-0.012	0.141^{***}	-0.021
Breakpoint/No Breakpoint	0.061^{***}	0.073^{***}	0.099^{***}	0.041^{***}
R-squared	0.027	0.030	0.145	0.029
adj. R-squared	0.026	0.030	0.143	0.028
N	12,096	21,024	4,320	12,960

$^{***}\ p < 0.001$ $^{**}\ p < 0.01$ $^{*}\ p < 0.05$

Table 6.3: Results of regression analyses of the relative MSE differences for the different types of time series. The regression analyses confirm the previous results and derived guidelines. Most importantly, depending on the time series type, choosing an appropriate transformation to address non-stationarity is found to be essential.

The results regarding the weighting indicate that there are only significant differences for trended and seasonal time series. For trended time series, learning weights positively influences the correction, however –as the estimates for the interaction terms indicate– only in case of no transformation or detrending. This result is in line with the previous discussion of the weighting approaches. In contrast, learning weights has a significant negative influence for seasonal time series.

Using LAD instead of OLS results in significant differences in favor of LAD only for seasonal time series. However, judging from the interaction term, this is only the case for time series with low to medium variability. A high variability in general reduced the performance of forecast correction method, except for stationary time series.

Lastly, the regression clearly indicates that considering breakpoints in any form has a significant negative influence on forecast correction, independently of the time series type.

Overall, the regressions confirm the previous discussions and the significance of the effects on which the derived guidelines are based. While the analyses up to this point demonstrated which model to use under which conditions, an additional interesting point is how the different models actually work, i.e., which parameters they estimate for the correction. This question is addressed next in an analysis and discussion of the parameter estimates.

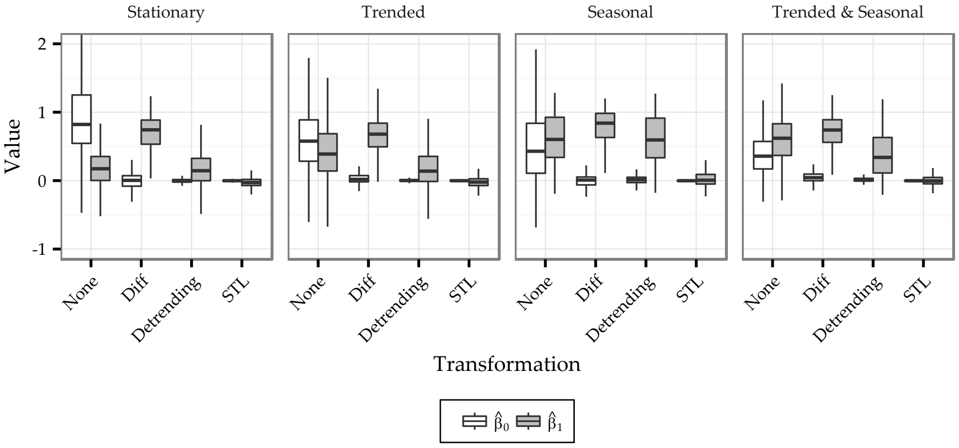


Figure 6.5: Distribution of slope and (scaled) intercept estimates for different data transformation methods and time series types. The estimates differ strongly between time series types and transformations. All parameters indicate substantial damping of the original judgmental forecasts.

6.2.5 Discussion of Parameter Estimates

As the previous analyses have shown, advantageous correction methods can be found for all time series types. Thus, the corresponding models seem to successfully identify linear biases that can be removed in order to improve accuracy. To gain further insights into the biases affecting forecast accuracy, the parameters estimated by the forecast correction models are now analyzed.

Figure 6.5 displays the distribution of the estimates of slope and intercept by time series type and by data transformation method. As the intercept estimates can vary by orders of magnitude, depending on the scale of the time series, the intercept estimates are scaled by the mean of the actual values in order to ensure comparability of the intercepts across time series.

Reconsidering that an unbiased forecast results in an estimated intercept $\hat{\beta}_0 = 0$ and slope $\hat{\beta}_1 = 1$, substantial deviations from unbiasedness can be noted for the results of the case study. The results are now discussed in detail for each data transformation approach.

First, without transformation, intercepts are very high for stationary time series whereas slopes are very low. Consequently, the original forecasts have only weak influence on the corrected forecast. Instead, a high intercept is estimated close to the mean of the time series, which is then the major part of the corrected forecast. Thus, for stationary time series, a very strong damping effect towards the mean of the time series can be observed. To a weaker extent, this effect can also be observed for trended as well as seasonal time series. In both cases, a substantial portion of the mean of the time series is used as a basis. However, in order to match the trend or seasonality of the time series, a stronger influence of the original forecast is required. This is even more pronounced for time series with trend and seasonality, where the original forecasts are only slightly damped.

Second, when applying differentiation to the time series, slopes are high but mostly substantially below one for all types of time series. Estimated intercepts are in contrast very low. Consequently, for differentiated data, the predicted changes are slightly damped, which is in line with the reasoning for the previous case without transformation.

Third, for detrending of the time series, intercepts for all types of time series are close to zero. Consequently, after adjusting for the trend, there are no large mean biases left. However, the predicted deviations from the trend component are weighted down very strongly for stationary and trended time series. Thus, a large regression bias exists and the judgmental forecasts exaggerate deviations from the systematic trend component. Only if a seasonal component exists (i.e. time series of type seasonal or trended and seasonal), a large portion of the deviations from the trend are included in the corrected forecast. This effect is explained by the seasonal component, which should –at least approximately– be included in the forecasts whereas it is not included in the trend component.

Lastly, when using STL to remove a trend as well as a seasonal component, the estimated intercepts are very close to zero, most likely as a result of the included detrending of the time series. However, in addition to this previously discussed effect, the estimated slopes are also very low. Accordingly, in the correction, only small deviations from the systematic trend and seasonal components are introduced and the expert forecast has little influence on the corrected forecast. As a consequence, the experts who produced the judgmental forecasts seem to have little skill in predicting deviations from the systematic components.

Overall, independently of the data transformation used, the findings indicate that the forecasts are –on average– damped by the forecast correction methods. A substantial part of the original judgmental forecast is replaced by a systematic component, such as the estimated intercept without preprocessing or the identified trend and/or seasonal components. While this result may be surprising at a first glance, comparable observations have been reported in the literature in other contexts. Amongst others, Eggleton (1982) and O'Connor et al. (1993) found that one of the main problems with judgmental forecasts is that forecasters often wrongly identify systematic patterns in the noise of a time series. As a result, forecasters tend to move their forecast in the direction of their expected patterns (resulting in too extreme values), which in turn results in reduced accuracy as forecasts vary too strongly. Lopes and Oden (1987), Andreassen (1988), and Harvey (1995) provide more studies showing that people tend to respond to randomness as if it was signal and found that accuracy decreases disproportionately with reduced predictability of a forecasting task.

By applying forecast correction, these effects are –at least partially– corrected by reducing the variability of the judgmental forecasts. This approach results in increased accuracy, as demonstrated in this section. Furthermore, the damping effect of the correction motivates the importance of the choice of an adequate transformation of the time series as damping is only reasonable after adjusting for systematic components in the time series.

Overall, the influence of the original judgmental forecast is in many cases smaller than could be expected and the corrected forecast mostly consists of a prediction of the systematic components of the time series. While this effect is indirectly achieved in forecast correction by introducing transformations of the time series, it can directly be achieved in forecast combination methods. Statistical time series forecasting methods aim at predicting systematic components with high accuracy. Thus a combination of forecasts, which is analyzed in the next section, can result in similar effects.

6.3 Results on Forecast Combination

Various forecast combination methods have been proposed in the literature that did not systematically outperform the SA of forecasts. In this section, the optimal and robust shrinkage approaches introduced in this work are evaluated and compared to a SA combination.

As a metric for the performance of the different forecast combination approaches, the MSE is again used as it is the metric optimized by weight estimation methods. To ensure comparability, the relative MSE difference to the corresponding SA combination is used as SA is the benchmark method to be outperformed in forecast combination. Negative values indicate that an alternative combination outperformed SA.

In a first step, the mean and the median of the relative MSE can be analyzed for different approaches. The significance of the difference of the mean from zero can be tested using a t-test.

As preliminary evaluations showed that the median of the relative MSE difference is always very close to zero, additional quantiles are analyzed. The quantiles allow an evaluation whether the accuracy increases outweigh the decreases, i.e., whether the distribution is skewed in favor or to the disadvantage of a combination. The skewness of the quantiles can be quantified by calculating the quantile skewness (Groeneveld, 1991) introduced by Bowley (1901), as defined in Equation 6.6. A positive quantile skewness indicates a skew in favor of SA whereas a negative value indicates that the alternative combination is beneficial.

$$QS_{\alpha} = \frac{(x_{1-\alpha} - x_{0.5}) - (x_{0.5} - x_{\alpha})}{x_{1-\alpha} - x_{\alpha}} \tag{6.6}$$

To test whether the skew is significantly different from zero, a bootstrap test with 5,000 replications is used.

Using these metrics, the forecast combination methods, established or proposed in this work, can be analyzed for different sets of forecasts included in the combination. First, results of the combination of the two model-based forecasts are analyzed. The combination of a model-based with the judgmental forecast are evaluated subsequently. Lastly, results are provided for a combination of all three available forecasts.

6.3.1 Two Model-Based Forecasts

Different model-based forecasts are likely to differ regarding the ability to predict specific aspects of time series. Thus, combinations of model-based forecasts can often be beneficial. Depending on the errors of the forecasts, a SA combination may be reasonable in many cases whereas other cases allow a stronger weighting of one forecast.

Table 6.4 presents the results for the combination of the two model-based forecasts for different methods (first column) and robustness parameters (second and third column, if applicable). Besides the mean relative MSE difference to a SA combination (fourth column), different quantiles (fifth to ninth column) as well as the skewness of the quantiles (last two columns) are presented.

A combination using OW is already slightly beneficial in comparison to SA regarding the outer quantiles of the relative MSE difference. The 10 % and 90 % quantiles indicate that improvements of 39 % are not completely negated by the decreases in accuracy of 32 %. This is confirmed by the significantly negative skewness of these quantiles. However, mean as well as median relative MSE differences are clearly positive.

In contrast to the OW combination, clearer advantages can be noted if optimal shrinkage or robust shrinkage is used. Especially for higher robustness parameters, the mean relative MSE difference is significantly negative and the quantiles are additionally significantly skewed in favor of the combination with robust shrinkage. For instance using robust shrinkage with 0.1 for both robustness parameters results in only minor decreases in accuracy of 0.2 % for the 75 % quantile

				Quantile					Skewness	
Method	r	v	Mean	0.1	0.25	0.5	0.75	0.9	0.25	0.1
OW			0.06^{***}	-0.39	-0.08	0	0.09	0.32	0.01	$-0.11^{\dagger}$
Opt. Shr.			0.00	-0.33	-0.06	0	0.04	0.20	$-0.20^{\dagger}$	$-0.25^{\dagger}$
Rob. Shr.	0.1	0.1	-0.03^{***}	-0.32	-0.05	0	0.02	0.14	$-0.36^{\dagger}$	$-0.40^{\dagger}$
		0.3	-0.04^{***}	-0.23	-0.00	0	0	0.06	$-1.00^{\dagger}$	$-0.61^{\dagger}$
		0.6	-0.02^{***}	-0.01	0	0	0	0	n. def.	$-1.00^{\dagger}$
		0.9	0.00	0	0	0	0	0	n. def.	n. def.
	0.3	0.1	-0.04^{***}	-0.30	-0.05	0	0.02	0.10	$-0.44^{\dagger}$	$-0.49^{\dagger}$
		0.3	-0.04^{***}	-0.23	-0.00	0	0	0.05	$-1.00^{\dagger}$	$-0.63^{\dagger}$
		0.6	-0.02^{***}	-0.01	0	0	0	0	n. def.	$-1.00^{\dagger}$
		0.9	0.00	0	0	0	0	0	n. def.	n. def.
	0.6	0.1	-0.04^{***}	-0.25	-0.03	0	0.01	0.07	$-0.52^{\dagger}$	$-0.55^{\dagger}$
		0.3	-0.04^{***}	-0.22	-0.00	0	0	0.05	$-1.00^{\dagger}$	$-0.66^{\dagger}$
		0.6	-0.02^{***}	-0.01	0	0	0	0	n. def.	$-1.00^{\dagger}$
		0.9	0.00	0	0	0	0	0	n. def.	n. def.
	0.9	0.1	-0.04^{***}	-0.21	-0.03	0	0.01	0.05	$-0.57^{\dagger}$	$-0.60^{\dagger}$
		0.3	-0.04^{***}	-0.19	-0.01	0	0	0.03	$-1.00^{\dagger}$	$-0.70^{\dagger}$
		0.6	-0.02^{***}	-0.01	0	0	0	0	n. def.	$-1.00^{\dagger}$
		0.9	0.00	0	0	0	0	0	n. def.	n. def.

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$ † $p < 0.05$ (bootstrap test)

Table 6.4: Mean, quantiles, and quantile skewness of the relative MSE difference between different combination methods and SA when combining DTES and ARIMA forecasts. Optimal shrinkage is already beneficial regarding the quantiles. Increasing robustness to a moderate level results in significant performance increases.

whereas the 25 % quantile corresponds to an improvement of 5 %. This effect is even more pronounced for the outer quantiles, where accuracy decreases of 14 % are more than compensated by improvements of 32 %. The mean relative MSE difference is in this case −3 %, which confirms the significant skew in favor of the robust shrinkage combination.

The mean relative MSE difference is optimized when at least one of the two robustness parameters is set to a value around 0.3. Increasing the robustness parameters further again decreases the advantage in comparison to SA as higher robustness requires stronger shrinkage, which in turn results in a high similarity to SA.

Thus, in a combination of model-based forecast, the optimal and robust shrinkage levels not only provide insights into theoretical aspects of forecast combination, but also result in significant advantages over the SA combination. While the optimal shrinkage level is already beneficial, increasing robustness to a moderate level further increases the performance advantage over SA.

The next analyses provide insights on the applicability when the combination not only involves model-based but also judgmental forecasts.

6.3.2 One Judgmental and One Model-Based Forecast

For the evaluation of combinations of judgmental and model-based forecasts, two combinations can be analyzed. First, results for the combination of judgmental and ARIMA forecasts are provided in Table 6.5. Second, the combination of judgmental and DTES forecasts is evaluated in Table 6.6.

For the combination of judgmental and ARIMA forecasts, neither a combination using OW nor using the proposed optimal shrinkage factor improves over the SA combination. The mean relative MSE difference is significantly positive for both methods and is even in the best case 8 %. The quantiles furthermore have a significant positive skew, indicating that cases with accuracy decreases in comparison to SA additionally significantly outweigh the cases where the accuracy improved.

However, increasing the robustness parameters has a substantial influence on the mean MSE difference as well as on the quantiles and their skewness. The differences vanish if the robustness parameter indicating robustness against error variance changes reaches $v = 0.6$. If v is further increased to 0.9, a very small and significant improvement over SA can be noticed, albeit only with $p < 0.05$. Thus, if judgmental and ARIMA forecasts are combined, a very high robustness is required, which in most cases results in shrinking the weights completely towards SA.

Results are similar but less pronounced for the combination of judgmental and DTES forecasts. While OW and optimal shrinkage significantly decrease accuracy in comparison to SA, some combinations using robust shrinkage result in significant improvements. For instance for $v = 0.6$, all relative MSE differences are negative and significantly different from zero. In addition to the negative mean of the MSE difference, the quantiles are skewed in favor of the robust shrinkage combination. However, increasing to robustness against error variance changes to much (to $v = 0.9$) again decreases the accuracy improvements. Furthermore, only increasing the robustness against changes of error correlation has only a small influence on the forecast combinations.

In summary, when combining one judgmental with one model-based forecast, choosing relatively high robustness parameters, especially for changes of error variances, are required. This strongly contrasts the previous results for the combination of two model-based forecasts, where a much lower degree of robustness was required to outperform SA. As these results are very different, an obvious question is how combining two model-based forecasts with the judgmental forecast affects the choice of the combination method. This aspect is analyzed next.

				Quantile					Skewness	
Method	r	v	Mean	0.1	0.25	0.5	0.75	0.9	0.25	0.1
OW			0.13 ***	−0.31	−0.09	0.02	0.23	0.63	0.31 †	0.30 †
Opt. Shr.			0.08 ***	−0.27	−0.06	0	0.13	0.46	0.37 †	0.25 †
Rob. Shr.	0.1	0.1	0.07 ***	−0.27	−0.05	0	0.11	0.42	0.34 †	0.21 †
		0.3	0.05 ***	−0.23	−0.02	0	0.05	0.35	0.39 †	0.19 †
		0.6	0.01	−0.01	0	0	0	0.01	n. def.	−0.31
		0.9	−0.00 *	0	0	0	0	0	n. def.	n. def.
	0.3	0.1	0.05 ***	−0.27	−0.05	0	0.09	0.38	0.27 †	0.16 †
		0.3	0.04 ***	−0.24	−0.02	0	0.04	0.34	0.34 †	0.17 †
		0.6	0.01	−0.01	0	0	0	0.01	n. def.	−0.31
		0.9	−0.00 *	0	0	0	0	0	n. def.	n. def.
	0.6	0.1	0.03 ***	−0.24	−0.05	0	0.07	0.32	0.12	0.14 †
		0.3	0.03 ***	−0.23	−0.02	0	0.03	0.3	0.16	0.13 †
		0.6	0.01	−0.01	0	0	0	0.01	n. def.	−0.34
		0.9	−0.00 *	0	0	0	0	0	n. def.	n. def.
	0.9	0.1	0.02 **	−0.22	−0.04	0	0.05	0.26	0.04	0.09
		0.3	0.02 **	−0.21	−0.02	0	0.03	0.24	0.03	0.07
		0.6	0.00	−0.01	0	0	0	0.01	n. def.	−0.32
		0.9	−0.00 *	0	0	0	0	0	n. def.	n. def.

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$ † $p < 0.05$ (bootstrap test)

Table 6.5: Mean, quantiles, and quantile skewness of the relative MSE difference between different combinations and SA when combining Exp and ARIMA. A very high degree of robustness is require to achieve a small advantage over SA.

				Quantile					Skewness	
Method	r	v	Mean	0.1	0.25	0.5	0.75	0.9	0.25	0.1
OW			0.20 ***	−0.39	−0.09	0.03	0.21	0.40	0.17 †	0.23 †
Opt. Shr.			0.10 ***	−0.36	−0.07	0.01	0.11	0.47	0.18 †	0.12
Rob. Shr.	0.1	0.1	0.06 ***	−0.33	−0.06	0	0.09	0.38	0.19 †	0.07
		0.3	0.03 *	−0.28	−0.01	0	0.02	0.20	0.20	−0.16
		0.6	−0.03 ***	−0.10	0	0	0	0	n. def.	−1.00 †
		0.9	−0.00 **	0	0	0	0	0	n. def.	n. def.
	0.3	0.1	0.03 **	−0.34	−0.06	0	0.06	0.28	0.06	−0.10
		0.3	0.02	−0.28	−0.01	0	0.02	0.19	0.08	−0.21 †
		0.6	−0.03 ***	−0.10	0	0	0	0	n. def.	−1.00 †
		0.9	−0.00 **	0	0	0	0	0	n. def.	n. def.
	0.6	0.1	0.00	−0.3	−0.05	0	0.04	0.19	−0.10	−0.23 †
		0.3	−0.00	−0.28	−0.02	0	0.01	0.15	−0.28	−0.30 †
		0.6	−0.03 ***	−0.10	0	0	0	0	n. def.	−1.00 †
		0.9	−0.00 **	0	0	0	0	0	n. def.	n. def.
	0.9	0.1	−0.02 *	−0.28	−0.05	0	0.03	0.15	−0.18	−0.31 †
		0.3	−0.02 **	−0.27	−0.02	0	0.01	0.11	−0.45 †	−0.41 †
		0.6	−0.03 ***	−0.10	0	0	0	0	n. def.	−1.00 †
		0.9	−0.00 **	0	0	0	0	0	n. def.	n. def.

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$ † $p < 0.05$ (bootstrap test)

Table 6.6: Mean, quantiles, and quantile skewness of the relative MSE difference between different combinations and SA when combining Exp and DTES. A considerable robustness especially against variance changes is required to outperform SA.

				Quantile					Skewness	
Method	r	v	Mean	0.1	0.25	0.5	0.75	0.9	0.25	0.1
OW			0.39***	−0.36	−0.13	0.04	0.34	1.09	0.28†	0.45†
Opt. Shr.			0.17***	−0.34	−0.11	0.01	0.17	0.70	0.18†	0.34†
Rob. Shr.	0.1	0.1	0.10***	−0.33	−0.10	0	0.13	0.55	0.12	0.24†
		0.3	0.07***	−0.29	−0.06	0	0.05	0.38	−0.08	0.13†
		0.6	0.02*	−0.16	−0.00	0	0	0.08	−1.00†	−0.35†
		0.9	−0.01**	−0.00	0	0	0	0	n. def.	−1.00†
	0.3	0.1	0.06***	−0.31	−0.09	0	0.07	0.35	−0.08	0.07
		0.3	0.05***	−0.29	−0.06	0	0.04	0.32	−0.17	0.06
		0.6	0.02*	−0.16	0	0	0	0.08	−1.00†	−0.35†
		0.9	−0.01**	−0.00	0	0	0	0	n. def.	−1.00†
	0.6	0.1	0.02	−0.27	−0.04	0	0.01	0.18	−0.50†	−0.19†
		0.3	0.02	−0.27	−0.04	0	0.01	0.18	−0.64†	−0.21†
		0.6	0.01	−0.17	0	0	0	0.07	n. def.	−0.42†
		0.9	−0.01**	−0.00	0	0	0	0	n. def.	−1.00†
	0.9	0.1	−0.00	−0.19	−0.01	0	0	0.09	−1.00†	−0.37†
		0.3	−0.00	−0.19	0	0	0	0.09	−1.00†	−0.37†
		0.6	0	−0.14	0	0	0	0.03	n. def.	−0.61†
		0.9	−0.01***	−0.00	0	0	0	0	n. def.	−1.00†

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$ † $p < 0.05$ (bootstrap test)

Table 6.7: Mean, quantiles, and quantile skewness of the relative MSE difference between different combinations and SA when combining all three forecasts. A very high degree of robustness is required to achieve a small advantage over SA.

6.3.3 One Judgmental and Two Model-Based Forecasts

The theoretical analyses and discussions in Chapter 4 showed that a combination of three forecasts can often be expected to outperform a combination of two forecasts, even when considering the additional shrinkage required to meet robustness requirements. As a higher number of forecasts included in a combination might require other combination mechanisms, Table 6.7 provides results of the combination of all three available forecasts.

Overall, the results are mostly similar to the results for the combination of judgmental and ARIMA forecasts. OW and the combination using optimal shrinkage significantly again decrease the accuracy of the combined forecast in comparison to SA. Considering the positive mean relative MSE difference and the positive skewness of the quantiles, which are both mostly significantly different from zero, SA is clearly superior to these two alternative combinations. However, using robust shrinkage with a high robustness level regarding changes of error variances ($v = 0.9$) results in small but significant improvement over SA. As for the combination of Exp and DTES, the robustness against changes of the error variances is of higher importance than the robustness against changes of error correlation.

The analyses of combinations with different numbers and choices of forecasts overall show that only the robust shrinkage combination with appropriate robustness levels is a reasonable alternative to SA. Using the optimal shrinkage level, which adjusts for the estimation uncertainty, was in most cases disadvantageous in comparison to SA.

As adjusting for the estimation uncertainty is not sufficient and a considerable degree of robustness is required in all combinations, substantial changes of the error characteristics (especially the error variances) can be suspected for the forecasts of the case study. This especially seems to be the case for combinations involving the judgmental forecast as the combination that did not include the judgmental forecast required the lowest robustness. In order to derive more insights into this aspect, structural changes in the errors of the forecasts of the case study are analyzed next.

6.3.4 Structural Change in Practice

Forecast combinations with weights that are based on estimation from past observations can, as extensively discussed in Chapter 4, perform worse than a SA combination if the error characteristics change over time in a disadvantageous way. OW as an extreme approach can be expected to be especially prone to changes.

In the empirical evaluation, the results on forecast combination clearly confirm that OW combination is seldom superior to SA combination. Likewise, using optimal shrinkage only resulted in slight improvements (regarding the quantile skewness) over SA for the combination of the two model-based forecasts. Thus, adjusting for the uncertainty resulting from the estimation from finite samples is not sufficient for a successful combination. This result can be explained by structural changes within the error covariance matrix, which might especially be strong for judgmental forecasts. This is also confirmed by the fact that using robust shrinkage with high robustness settings did result in accuracy improvements in most cases.

Unfortunately, the true error covariance matrices are unknown in practice (especially for observations for one point of time) and can therefore not be analyzed for changes. As an alternative, rolling error correlations and error variances are estimated with a window of one year for the three years for which the judgmental as well as the model-based forecasts are available. The last estimated value in the rolling analysis is most likely to be closest the true value in the evaluation sample. Consequently, this value is used as a basis for comparing and analyzing the error correlations and error variances over time.

Figure 6.6 presents the differences between the estimated rolling error correlation or ratio of error variances and the last corresponding estimate for different

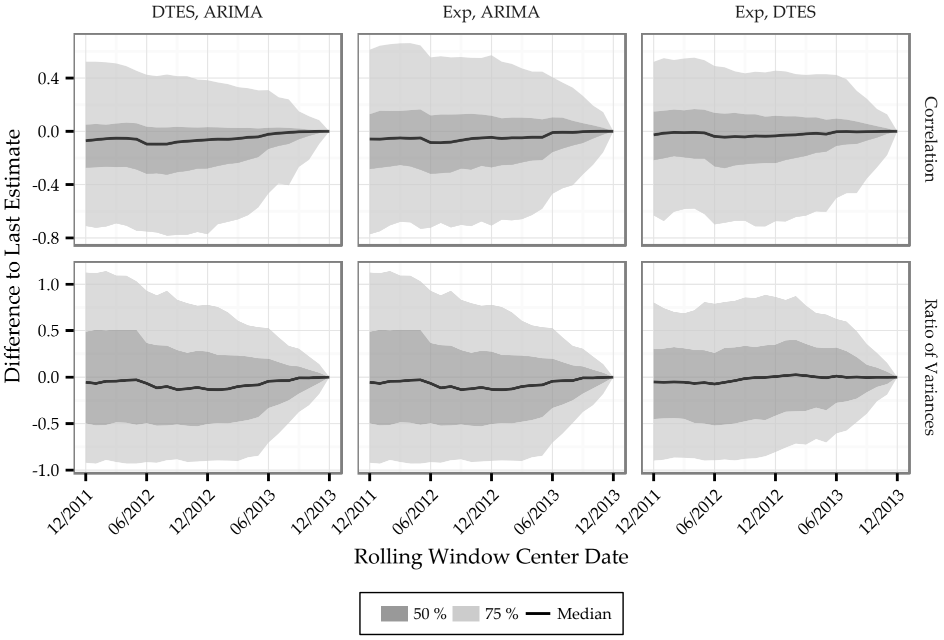


Figure 6.6: Distribution of the difference of rolling error characteristic estimates and the last estimate. Results for the error correlation as well as for the ratio of error variances are presented for different pairs of forecasts. A high uncertainty over time exists for all combinations.

pairs of forecasts. Ratios of error variances are used since, on the one hand, error variances differ between time series and, on the other hand, weights reflect the relative instead of the absolute performance of forecasts. The solid lines indicate the median of the difference between estimates whereas the dark gray (light gray) areas contain 50 % (75 %) of all observations per center date of the rolling window.

First, regarding error correlations (upper plot), the median fluctuates around zero for all pairs of forecasts, indicating that the overall distribution of error variances is centered on the same values. However, the quantiles of the difference increase for earlier error correlations (i.e. for earlier center dates of the rolling window). This result gives a first indication that the error correlations might systematically change over time. Furthermore, there is a substantial uncertainty regarding the error correlation over the different rolling window estimates. These results are very similar for all pairs of forecasts.

Second, regarding error variances, the difference between estimates of the ratio of error variances and the last estimate fluctuates, similarly to the error correlation, around zero. This again indicates that the estimates are overall centered on the same values while the uncertainty is high, the ratio of error variances can

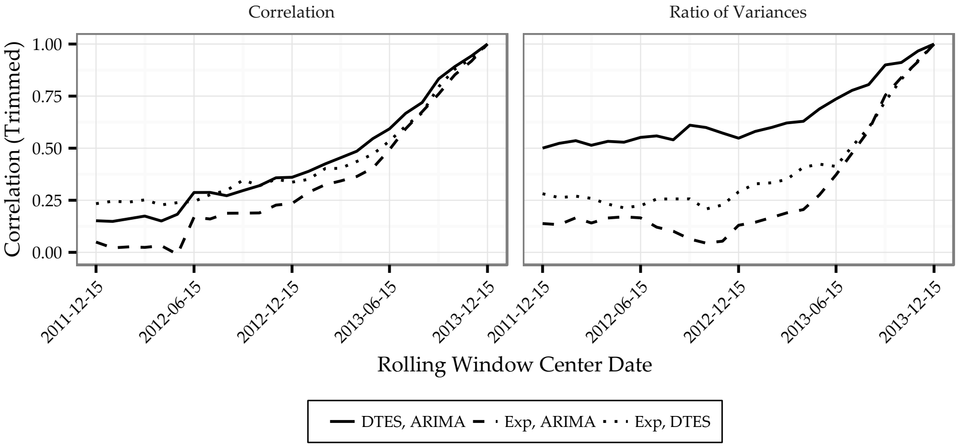


Figure 6.7: Correlation between different rolling estimates and the last rolling estimate as a proxy for the information content of older estimates regarding newer estimates. The error characteristics of the ARIMA, DTES pair has clearly highest predictability while Exp, ARIMA has the lowest.

change by ±1. The development of the ratio of error variances over time is very similar between the combinations of DTES and ARIMA on the one hand and the judgmental forecast and ARIMA on the other hand. Only the combination of the judgmental forecast and DTES differs slightly.

In summary, the analyses illustrate that there is considerable uncertainty regarding error variances and error correlation and that the overall distribution does not shift as the median is more or less constant. It can however not be concluded whether individual estimates also fluctuate around the last value or whether there are systematic changes. This aspect is addressed in Figure 6.7, where correlations (slightly trimmed at the 1 % and 99 % quantile) of the rolling estimates with the last rolling estimate are displayed. The correlation is used as a proxy for the information content of older estimates regarding newer estimates.

If only random fluctuations occurred per time series, the correlations could be expected to be relatively high, albeit decreasing with earlier estimates. However, the analysis shows that the correlation is in many cases very low, especially for early estimates.

Regarding the error correlation estimates, the correlation is close to or approximately zero, especially for early estimates for the combination of judgmental and ARIMA forecasts. The correlations then increase over time in a near-linear way. This clearly indicates that early estimates contain little information about later estimates and that systematic changes on the time series level occur frequently. The effect is most pronounced for the judgmental forecast and ARIMA where all correlations increase later and are at a lower level than those of the other two combinations.

Regarding the ratio of error variances, a similar effect can be noted except for the combination of DTES and ARIMA forecasts. The correlation of the ratio of error variances is in this case is substantially higher for all observations and already approximately 0.5 for the first estimate. The correlation again increases latest for the combination of judgmental and ARIMA forecasts.

Thus, although error variances and error correlations change for all forecasts, the changes are more or less systematic for different combinations. While the changes for DTES and ARIMA consist of more random fluctuations, changes are more systematic for the other pairs, especially the judgmental and ARIMA forecasts.

The observed changes in summary match the robustness parameters required for successful combination in the previous analyses very well. The combination of ARIMA and DTES required lowest robustness parameters and changes are found to be mostly random fluctuations. In contrast, the other pairs required stronger shrinkage for robustness and early estimates are found to contain little information in these cases. The observed structural changes thus explain the differing robustness required for different combinations.

Overall, the novel shrinkage approaches introduced in this work showed interesting results in the case study. Only controlling for estimation uncertainty by using optimal shrinkage only resulted in improvements over SA for one combination. In contrast, robust shrinkage resulted in improvements in most cases when an appropriate level of robustness was used. The required robustness degree depends on the forecasts included in the combination. Although changes are likely to occur for all forecasts, the changes differ between forecasts. While some changes can be explained by random fluctuations, other changes seem to be systematic. Finding an adequate degree of robustness of the combination models is thus an important basis for a successful combination of forecasts.

Up to this point, both forecast correction as well as combination models were found to be mostly beneficial in the previous analyses. As a consequence, the next section provides a comparison of the performance of the two approaches in order to derive comprehensive guidelines.

6.4 Comparison of Correction and Combination

In the previous sections, forecast correction and combination were analyzed separately. Main results were that appropriately treating non-stationarity of time series is of importance in forecast correction and that the novel robust shrinkage-based forecast combination method is of advantage if adequate robustness levels are chosen.

However, the previous analyses do not allow deriving comprehensive guidelines that include when to choose either forecast correction or forecast combination. For this purpose, Table 6.8 provides the median relative MSE difference to the judgmental forecast for the different approaches and by time series type. Additionally, ranks of the approaches per time series type are provided in parenthesis. From the various available methods in forecast correction, the methods with OLS estimation, no breakpoint and equal weighting are considered. For the forecast combination methods with robust shrinkage, different robustness requirements are used, depending on the forecasts included in the combination. Based on the results of the previous section, $r = v = 0.3$ is used for the combination of DTES and ARIMA forecasts as robustness requirements were found to be lowest in this case. For the combination of judgmental and DTES forecasts, $r = v = 0.6$ is used. As combinations including ARIMA as well as judgmental forecasts required the highest robustness, $r = v = 0.9$ is used in these cases. The robustness parameter used for the robust shrinkage combination is additionally stated in the table in parentheses behind the robust shrinkage method.

In case of stationary time series, forecast correction without transformation performs best amongst the methods under study. While it is clear from the previous analyses that using no data transformation results in forecast correction models with the best performance in case of stationarity, the analysis now shows that all combinations of forecasts are outperformed. Thus, the model-based forecast in these cases seem to add no additional information that is not yet considered by the forecast correction method. This finding is even clearer when considering that a no-change forecast is often relatively accurate for stationary time series. Thus, a combination of the judgmental with a model-based forecast results in a damping effect similar to the one in forecast correction. Amongst the different forecast combination approaches, a combination of DTES and the judgmental forecast is most beneficial. Comparing the different combination methods shows that SA performs best and only slightly outperforms the corresponding robust shrinkage combination for these two forecasts.

In contrast to the result for stationary time series, forecast correction models do not outperform forecast combinations for trended, seasonal, as well as trended and seasonal time series. For trended time series, a combination of DTES and the judgmental forecast is most accurate. As DTES is a forecasting model focusing on trends (that are additionally damped), it is not surprising that it outperforms combinations involving the more general ARIMA forecasts. Amongst the differ-

		Time Series Type			
Approach	Method	Stationary	Trended	Seasonal	Trended & Seasonal
Correction	No Transformation	−0.24 (1)	0.01 (19)	0.03 (12)	−0.00 (20)
	Diff	−0.02 (20)	−0.08 (7)	0.03 (11)	−0.07 (16)
	Detrending	−0.17 (6)	−0.06 (10)	0.03 (10)	−0.03 (19)
	STL	−0.03 (19)	0.02 (20)	0.09 (16)	−0.20 (10)
Combination	DTES, ARIMA / SA	−0.15 (10)	−0.04 (14)	0.91 (20)	−0.21 (9)
	DTES, ARIMA / OW	−0.14 (12)	−0.05 (13)	0.06 (15)	−0.24 (5)
	DTES, ARIMA / Opt. Shr.	−0.14 (11)	−0.06 (12)	0.05 (14)	−0.24 (6)
	DTES, ARIMA / Rob. Shr. (0.3)	−0.15 (9)	−0.06 (11)	0.05 (13)	−0.24 (6)
	Exp, ARIMA / SA	−0.15 (7)	−0.11 (5)	−0.16 (1)	−0.30 (2)
	Exp, ARIMA / OW	−0.05 (18)	−0.01 (18)	−0.05 (6)	−0.17 (13)
	Exp, ARIMA / Opt. Shr.	−0.09 (15)	−0.04 (16)	−0.07 (5)	−0.23 (8)
	Exp, ARIMA / Rob. Shr. (0.9)	−0.15 (7)	−0.11 (5)	−0.16 (1)	−0.30 (1)
	Exp, DTES / SA	−0.22 (2)	−0.16 (1)	0.85 (19)	−0.19 (11)
	Exp, DTES / OW	−0.08 (16)	−0.04 (15)	−0.04 (7)	−0.03 (18)
	Exp, DTES / Opt. Shr.	−0.14 (13)	−0.08 (8)	−0.03 (8)	−0.06 (17)
	Exp, DTES / Rob. Shr. (0.6)	−0.21 (3)	−0.16 (2)	−0.02 (9)	−0.17 (14)
	All / SA	−0.19 (4)	−0.13 (3)	0.40 (18)	−0.27 (3)
	All / OW	−0.07 (17)	−0.03 (17)	−0.10 (4)	−0.14 (15)
	All / Opt. Shr.	−0.10 (14)	−0.08 (9)	−0.12 (3)	−0.19 (12)
	All / Rob. Shr. (0.9)	−0.19 (4)	−0.13 (3)	0.26 (17)	−0.27 (4)

Table 6.8: Comparison of the median relative MSE differences between different forecast correction or combination methods and the judgmental forecast. For stationary time series, the standard forecast correction method performs best. For other time series, different combinations are most successful. Combinations with robust shrinkage often perform equal to or better than SA combinations.

ent combination methods, SA again slightly outperforms the robust shrinkage combination.

For seasonal as well as trended and seasonal time series, the combination of judgmental and ARIMA forecasts performed best. SA and robust shrinkage combination in both cases performed best; results are identical for seasonal time series (because of the high robustness requirement) and robust shrinkage marginally outperforms SA for trended and seasonal time series.

Thus, two results can be noted at a first glance. First, all of the most beneficial methods included the judgmental forecast and a statistical model. While correction performed best for stationary time series, the best combinations for the other types of time series always included the judgmental forecast. Second, the robust shrinkage combination performs very similar to the SA combination if the best set of forecasts is chosen for a time series type. This result however differs for other forecasts included in the combination.

Interesting results can for instance be noted for the seasonal time series. First of all, DTES has a relatively low performance for seasonal time series. This result can be derived from the fact that all SA combinations involving DTES substantially increase the errors over the errors of the judgmental forecast (up to 91 % for the combination of DTES and ARIMA). In these cases, a weighting different from equal weights is of high importance in order to reduce the weight of the forecast with low accuracy. Thus, the alternative approaches, especially the ones with optimal and robust shrinkage, improve performance in these cases. Similar results can be noted for combinations of DTES and ARIMA forecasts and trended or trended and seasonal time series. As a consequence, the shrinkage-based methods can be recommended in cases where the accuracy of the forecasts differs substantially.

The robust shrinkage level introduced in this work overall performs very similar to SA as high robustness parameters are used, which often result in a complete shrinkage towards SA. However, in cases where SA is not very well suited as the accuracy of the forecasts differs strongly, the robust shrinkage combinations offer additional benefits. Thus, using robust shrinkage with adequate robustness parameters instead of SA is overall a promising approach, which only substantially deviates from SA for combinations where a stronger weighting is likely to be reasonable.

Overall, forecast correction methods are very strong for stationary time series. While correction methods offer benefits for all other except seasonal time series, forecast combination methods perform substantially better in these cases. When using forecast combination, an SA combination is –as can be expected from the literature– often a good or the best choice. However, the proposed robust shrinkage combination is also a viable alternative as it performs better than SA if forecast accuracies differ substantially and very similar in other cases.

6.5 Conclusions and Limitations

While the previous chapters focused on understanding and improving the robustness of forecast correction and combination methods, this chapter evaluated whether the analytical results can be transferred to applications in practice.

The analyses showed that at least some of the extensions to forecast correction models, especially the ones aiming at ensuring robustness against non-stationarity, can increase the performance of forecast correction methods.

In forecast combination, the analyses confirmed that structural change of error characteristics is likely to be one of the main reasons for the success of simple averaging in forecast combination. Only for the combination of two model-based forecasts, which were found to be most stable regarding error characteristics, adjusting the weights to consider estimation uncertainty increased performance in comparison to the simple average. Thus, structural changes were found to be too strong in other cases, which consequently required a higher degree of robustness against changes in a combination. Especially if both the judgmental and the ARIMA forecast are involved in a combination, structural changes can be very large and require a very strong shrinkage. This confirms the results of the theoretical analyses that a complete shrinkage, i.e., using the simple average, is required if the error characteristics are too unstable. The robust shrinkage combination approach was overall found to be a successful approach when adequate robustness levels are used.

Although using forecast correction and combination methods often improves the overall accuracy, automatic correction or combination of forecasts bears the risk of changing originally accurate forecasts in a disadvantageous way. For instance, an expert might have considered knowledge on future events not derivable from past time series and error histories by statistical means. The results on forecast correction and combination (see for instance Figure 6.5) clearly show that strongly increased errors occur in some cases.

This issue is most likely hard to address using statistical means. If methods are designed towards maximum conservativeness, not only the disadvantageous but also the beneficial changes increasingly vanish. Thus, in order to mainly prevent strongly disadvantageous changes, an additional interaction with the human expert who produced the judgmental forecast is required. Critical cases, which can for instance be identified on the basis of the difference between the changed and the original forecast, can be presented to the forecaster in order to differentiate between cases that would result in strong accuracy improvements or declines. A solid decision basis is required for this task since the expert likely requires a comprehensible explanation of why and how a statistical model suggested changes to the original forecast for the decision. Blanc and Setzer (2015a) proposed a design of a forecast support system that aims at supporting this interaction.

The empirical evaluation is subject to several limitations. Most importantly, the data set used in the evaluation stems from one company and one application. Thus, the generalizability of the results might be limited. However, the time series and forecasts used in the case study are from different subsidiaries of the sample company, which are very different in terms of business characteristics and forecasting processes. As a consequence, it is likely that the time series are reasonable representatives for time series from real-world applications. As the forecasts are produced by different forecasters with various cultural backgrounds, biases can be expected to vary strongly across forecasts.

Furthermore, as a result of the design of the forecasting process in the sample company, only one judgmental forecast is available per time series. For this reason, only combinations of different model-based or of model-based and the judgmental forecasts could be analyzed in the evaluation. However, as including one judgmental forecast already results in high robustness requirements and thus very often using the simple average for the combination, introducing additional judgmental forecasts would most likely only allow a simple average combination.

As the time series are real-world time series, which additionally include data from the economic crisis that began 2007, various effects not considered in this work might be present. These effects might cause some of the results found in this chapter or might make other effects disappear. However, thorough analyses of the data, including analyses by experts at the sample company, were conducted to ensure a high degree of reliability of the data.

The filtering of the time series used to derive the data set of the empirical evaluation involved excluding short time series as well as time series with too many zero values. Thus, only recommendations for time series with sufficient length as well as a sufficient number of non-zero values can be derived from the analyses.

The empirical evaluation was furthermore limited to a small set of forecast correction and combination methods, which were discussed in this work. Although various approaches exist, especially in forecast combination, only few methods have been included in the evaluation for reasons of complexity. However, SA was included in the evaluation, which is the most important benchmark from a theoretical as well as from a practical point of view.

The discussions regarding structural changes in this chapter are based on the assumption that the chosen data analysis methods, especially the correlation analysis, are able to detect structural changes in the error characteristics. As no standard method exists for analyzing these aspects, a new approach was developed that is likely to detect the effects of interest. As the approach is not an established one, the effects that were found in the analysis might be different from the ones assumed and discussed in the analysis. However, the results of the discussion fit the other results of this chapter very well as they provide comprehensible explanations.

Overall, notwithstanding these limitations, the empirical evaluation allowed a variety of insights into the robust application of forecast correction and combination in practice. These results and the derived guidelines matched the theoretical analyses and discussions very well and reveal promising starting points for future work, as is discussed in the next part of this work.

Part IV

Finale

Chapter 7

Conclusions and Outlook

JUDGMENTAL forecasts are widely used in practice as human experts can include contextual information into the forecasts. However, judgmental forecasts are regularly found to be inefficient as a result of cognitive biases and heuristics. An integration with quantitative statistical methods aims at increasing the accuracy of judgmental forecasts by mitigating these issues.

In this work, the robustness of forecast correction and combination, two established integration methods, was studied. Forecast correction aims at identifying systematic biases in past judgmental forecasts, which can then be removed from new forecasts. In contrast, forecast combination does not alter the original judgmental forecasts but uses a linear combination with alternative model-based ones. Applying the approaches requires estimating parameters from past data. In forecast correction, the parameters quantify the biases found in past forecasts whereas the weights of the forecasts are parameters that have to be chosen or estimated in forecast combination.

The statistical learning theory indicates that using estimated parameters in a model influences the error of the model outcomes on unknown data. While complex models can be expected to have low systematic errors (low bias component), the estimation uncertainty can result in additional errors (high variance component). In contrast, simple models have few parameters and thus small errors resulting from estimation uncertainty but can be expected to introduce systematic errors (high bias and low variance component). As a result of this trade-off, models can be expected to be influenced differently by small training samples, one of the key aspects driving estimation uncertainty. Another issue introducing uncertainty in practice is structural change, which results in systematic differences between past observations and future unknown data.

Analyzing and understanding these influences is a key requirement for developing models that can be robustly applied in practical applications.

7.1 Contribution

This work aimed at understanding the theoretical properties and trade-offs of forecast correction and combination methods regarding robustness against estimation uncertainty as well as structural changes. The derived insights were used to develop novel mechanisms which aim at transferring the theoretical results to applications in practice in order to achieve a more robust performance on unknown data.

Although forecast correction and combination are established approaches, there is little previous research analyzing the robustness of the methods. Most importantly, there is no previous work that explicitly aims at understanding and solving the involved trade-offs. In order to summarize the contributions of this work, the individual contributions are discussed below on the basis of the research questions introduced in Chapter 1.

Forecast Correction – Training Sample Size: Although small training samples are known to result in increasingly unstable parameter estimates, which in turn increase errors on unknown data, this influence has not yet been analyzed in forecast correction methods. The theoretical analyses in this work indicated that, at least under the assumption of multivariate normality, the training sample size required for the corrected forecast to outperform the original judgmental forecast is relatively small. Training samples available in practice might only be too small if the removable biases are weak, i.e., if the judgmental forecasts are close to unbiasedness. The empirical evaluation showed that forecast correction overall results in improvements if other aspects, especially non-stationarity, are addressed adequately. Thus, considerable removable biases exist and the training sample sizes available in practice are not an issue for forecast correction.

Forecast Correction – Structural Breaks: Biases in judgmental forecasts depend on the human experts who produce the forecasts. The biases can change over time as a consequence. For instance, biases can decrease over time due to learning effects. Or, alternatively, biases can change completely because of a change of the forecaster. Changing biases also influence forecast correction models that use past data to identify biases that are then removed from future forecasts. If a bias change occurs, outdated biases are removed from future forecasts, which can result in the new biases not being optimally removed or even in introducing additional biases. Despite the high likelihood of structural changes in practice, the influence of these changes has not yet been researched. The analyses in this work identified that as long as the strength of the bias is not reduced too much towards unbiasedness, forecast correction methods can be expected to be largely robust against structural changes. In other words, the corrected forecast can be expected to be more accurate than the original forecast expect for relatively strong changes of the biases. However, although the corrected forecast is likely to be more accurate than the original one, structural breaks still influence the accuracy of the corrected forecast and the results of the correction are suboptimal.

In order to address structural changes, the existing approach using exponential weighting of past observations or the new approach proposed in this work, which explicitly treats detected structural breaks, can be used. Although including structural changes directly or indirectly in the weight estimation procedure could be expected to be beneficial, the empirical evaluation with real-world data in this work showed that including structural breaks does not improve the accuracy of corrected forecasts. Thus, although structural changes influence the accuracy of corrected forecasts, the additional instability and uncertainty in the correction model resulting from treating potential structural breaks likely outweighs possible accuracy gains.

Forecast Correction – Non-Stationarity: Previous studies mostly applied forecast correction methods directly to the time series data, independently of the characteristics of the time series. In this work, non-stationarity of time series, for instance because of a trend or seasonality, was shown to be a relevant issue in the theoretical analyses. In case of non-stationarity of time series, the biases that are considered by linear forecast correction models are increasingly undetectable. Removing biases is consequently increasingly impossible for non-stationarity and even a strongly biased forecast can be detected as unbiased. However, if forecast correction is nevertheless applied, the parameter estimation introduces additional uncertainty, and consequently errors, while no biases can be removed. Thus, the overall error of the corrected forecast can easily increase over the original forecast. This result is confirmed by the empirical evaluation. Applying forecast correction without addressing non-stationarity of time series did not result in improved forecast accuracy for trended time series as well as for time series with trend and seasonality. Especially for these time series, a detrending or deseasonalization was found to be essential for increasing the accuracy of judgmental forecasts. In summary, non-stationarity of time series is a key issue that has to be addressed for a successful application of forecast correction in practice.

Forecast Combination – Bias–Variance Trade-Off: The bias–variance tradeoff from statistical learning theory allows understanding how the performance of a statistical model is related to the expected fit of the model to the data on the one hand and to estimation uncertainty on the other hand. The trade-off is based upon a decomposition of the error on unknown data into two error components quantifying these aspects. Although the decomposition provides a useful framework for understanding the performance of statistical models with estimated parameters, no decomposition of the error of a combined forecast has been introduced in the literature. In this work, the decomposition of the combined error variance, including the required and previously unknown general case of the sampling distribution of the weight estimates, has been derived for the case of weights that minimize the in-sample error variance shrinked towards equal weights.

The formulation of the decomposed expected error variance of a forecast combination allowed analytically deriving the minimal training sample size required to outperform an alternative combination. The threshold can for instance be used to determine whether a combination with a specific shrinkage can be expected to outperform an equal weights combination. The discussion of the threshold value using forecast errors from the M3 Competition revealed that many combinations could be expected to outperform an equal weights combination for relative small sample sizes, in many cases below 20.

Furthermore, given the error covariance matrix of the forecasts, a shrinkage level was derived that optimally solves the bias–variance trade-off involved in forecast combination by balancing potential accuracy gains and errors result from estimation uncertainty. The discussion using the forecast errors from the M3 Competition illustrated that, using this shrinkage level, the expected combined error variance can be expected to continuously decrease with increasing training sample size and number of forecasts in a combination. While the former result is expected, the latter one contradicts existing guidelines on forecast combination, which suggest including only a limited set of forecasts in the combination. Furthermore, in the empirical evaluation, the optimal shrinkage method in most cases performed better than in-sample optimal weights but did not outperform an equal weights combination. Slight advantages over equal weighting were only found for combinations of model-based forecasts. Thus, in real-world applications, only a part of the error of a combined forecast can be attributed to estimation uncertainty and additional effects influence the errors.

Forecast Combination – Structural Changes: Similar to forecast correction, structural changes in the error characteristics can influence forecast combination methods. If error patterns differ too strongly between past observations and future errors, the combination of forecasts does not reduce error levels and can even result in strongly increased errors. Although structural changes are especially likely to occur in the combination of judgmental or judgmental and model-based forecasts, this aspect has received scant attention in the literature.

In order to assess the relevance of structural breaks for forecast combination models, critical changes have been derived, which quantify how strongly error characteristics are allowed to change for a combination to still outperform an alternative one such as an equal weights combination. The discussion of the thresholds showed that combinations of more than three to four forecasts are highly prone to changes of the error characteristics. For instance small changes of one error variance can result in combined error variances increased by a multiple over the error variance expected without a change. Thus, structural changes can explain why combinations of too many forecasts are not beneficial and often not recommended in the literature.

As the theoretical analyses clearly show the strong influence of structural changes, a potential approach to improve the performance of forecast combination is increasing the robustness against structural changes. Based upon the derived critical changes for a combination with a certain shrinkage level and a robustness requirement in terms of maximum changes, a robust shrinkage level was derived. This robust shrinkage level ensures that the combined forecast can be expected to perform as least as good as an alternative combination, such as an equal weights combination, as long as changes do not exceed the robustness requirements. The discussion using the forecast errors of the M3 Competition revealed that combining three to five forecasts is most promising when aiming at robustness by using the robust shrinkage level. While this finding explains existing guidelines, the analyses also showed that using a random set of forecasts is often a better choice than selecting the most accurate ones. This aspect can mainly be attributed to the high correlation of the most accurate forecasts, which results in lower diversity and thus lower robustness.

In the empirical evaluation, the robust shrinkage level was shown to improve over optimal shrinkage, which only considers estimation uncertainty. However, if combinations involve a judgmental forecast, error characteristics were found to be very unstable over time and a high robustness level is required, which in turn often results in using the simple average of forecasts. If adequate robustness levels are used, the robust shrinkage performs very similar to a simple average combination in many cases but performs better if the accuracy differs strongly between forecasts. Thus, the theoretical analyses provided important insights into the robustness against structural changes as well as means to addressing these changes. In order to transfer the theoretical results into practice, the choice of the robustness requirement is a key aspect.

Overall, the theoretical analyses in this work allowed deep insights into the robustness and performance of forecast correction and combination methods. The results on forecast correction can be easily transferred into practice and improve the real-world performance of forecast correction methods and thus the efficiency and efficacy of corporate processes relying on judgmental forecasts. Likewise, the introduced optimal and robust shrinkage levels can be implemented in practice for the combination of model-based and judgmental forecasts while adjusting for the differing likelihood and strength of structural changes by setting adequate robustness requirements for the combination.

Although a detailed understanding of the aspects influencing the robustness of the methods has been developed in this work, a variety of starting points for future work exist, which especially aim at improving the transfer of the theoretical results into practice. A selected set of promising directions of future work is presented in the next section.

7.2 Future Work

Developing a theoretical understanding of the various aspects influencing forecast correction and combination methods has in this work been demonstrated to be an important basis for improving existing approaches as well as for developing new ones. The proposed directions of future research thus focus on starting points for further theoretical insights as well as for approaches that better transfer the insights derived in this work into practice. First, future work on correction and subsequently on forecast combination is discussed. Lastly, future work beyond the two methods is proposed.

While the theoretical analysis of the robustness of forecast correction methods against structural changes focused on changes of only one parameter, simultaneous changes of the biases are likely to occur in practice. For instance, if an expert producing judgmental forecasts changes, not only the error variance but also the correlation between forecasts and realizations is likely to change. Thus, analyzing simultaneous bias changes can provide additional insights into the robustness. Furthermore, analyses of the biases in real-word forecasts could reveal how biases change in practice and how strong occurring changes are, which would provide additional insights in combination with the results of the theoretical analyses.

As detecting and explicitly considering structural breaks in the estimation of the forecast correction model did not result in improved accuracy of the corrected forecasts, novel approaches to addressing changes could be researched. Dynamic linear models, which allow estimating time-varying parameters (see for instance Harrison and West (1999)) would be an interesting approach that allows considering structural changes, especially slow bias changes, in a more flexible way. As an alternative, the so-called indicator saturation estimation can be used, which allows a very flexible detection of structural changes on the basis of predefined change pattern (see for instance Pretis et al. (2016) on how the approach can be used in a time series context).

Using more flexible estimation methods can be seen as a part of the more general problem that the detection of structural breaks and the transformation of the time series are treated as tasks that are independent from the estimation of the parameters of the correction model. However, there are strong dependencies between the individual steps. Identified breakpoints as well as the chosen transformation of the time series strongly influence the identifiable biases and the estimated parameters. Thus, all steps should in principle be geared towards maximizing the accuracy of the corrected forecast. However, separating the tasks into independent steps does not guarantee that the overall approach is well suited for the correction. For instance the breakpoint detection aims at identifying structural breaks that separate the data as cleanly as possible. However, an identified breakpoint break must not necessarily be of importance for the forecast correction as it might for instance be relatively weak. Likewise, the data transformation methods only aim at finding a preprocessing of the time series that ensures stationarity as good as possible. The chosen transformation can however not be guaranteed to result in stationary time series that can be corrected in a beneficial way. Overall, an integration of the different steps into one model is required in order to enable finding a model that improves the bias–variance trade-off when considering all of the relevant estimation and preprocessing steps.

In forecast combination, deriving novel means of estimating and shrinking weights are a promising area of research. The shrinkage applied in this work shrinked all weight estimates towards equal weights by the same percentage. While this already affects extreme weights more strongly than weaker ones, alternative shrinkage methods are of interest. For instance a non-linear shrinkage can be used, which does not shrink all forecasts in the same way but shrinks the weights of precisely those forecasts most strongly that are either unimportant for the combination (and thus only introduce errors because of estimation uncertainty) or that make the combination prone to structural changes. For instance, in this work, using shrinked weights outperformed the simple average in the combination of different model-based forecasts for an adequate robustness level. However, when a judgmental forecast was additionally included, a much higher degree of robustness was required to achieve performance comparable to the simple average. Thus, including the judgmental forecasts, which were shown to have stronger structural changes, was a clear issue for using shrinked weights. In cases such as this one, strongly shrinking the forecast that is likely to have structural changes except for cases where it is highly important might be beneficial. The idea of shrinking the parameter estimates differently is established in regression theory, where for instance lasso regression (Tibshirani, 1996) is used. Non-linear shrinkage furthermore enables shrinking a forecast out of the combination, i.e., reducing its weight to zero. An interesting question is how strongly this effect would be used in non-linear shrinkage and whether it could also explain the existing guidelines regarding the number of forecasts included in a combination.

In this work, shrinkage was not only used to reduce the model instability resulting from estimation errors, but also to achieve robustness against structural changes. Structural changes were analyzed in terms of critical changes, which were then used to determine weights that are, to a definable extent, robust against changes. However, the critical changes only considered changes of one error characteristic. Simultaneous changes, which can regularly occur in practice, were not considered. The model in this work could, in principle, be extended to consider multiple changes. However, the robust shrinkage level would then have to consider the worst case combination of changes. This would in turn result in a very conservative shrinkage factor and thus using equal weights in practically all cases. As a consequence, new approaches for determining a robust shrinkage level are of great interest. One approach could be based on resampling methods, which are often used in the literature to assess the estimation uncertainty. If error covariance estimates, or the resulting weight estimates, are very uncertain across resampling iterations, a stronger shrinkage must be used for robustness. Even more advanced approaches could aim at taking changes over time into account by using a weighted resampling method or by deriving reasonable robustness requirements or a robust shrinkage level from the observed development of the error characteristics in the past. This approach could address one of the key issues identified in this work, correctly setting the robustness requirements that are used for calculating the robust shrinkage level.

While forecast correction and combination are two established approaches to improving forecast accuracy, the two approaches are mostly implemented and applied separately, as has been done in this work. The only study analyzing the combination of both approaches was conducted by Goodwin (2000), who first corrected the judgmental forecast and then combined the corrected forecast with a model-based one. Although both individual approaches were found to be beneficial, the combination of the approaches did not result in additional significant improvements of the forecast accuracy. As the discussions in the empirical evaluation in this work revealed, there is a similarity between forecast correction and combination. The parameters estimated by the forecast correction models indicated that the judgmental forecasts are substantially damped towards the systematic component in the time series (or a constant value in case of stationarity). As model-based forecasts aim at modeling and predicting this systematic component, a combination of the judgmental forecast with a model-based one also often results in a shift towards the systematic component. This similarity is likely to result in the low additional benefit of using both methods as most of the improvements are already realized by applying one of the two methods. While the qualitative relationship is largely clear, the exact nature of the similarity is unknown. Furthermore, applying both methods might be beneficial in some cases and not be reasonable in others. Thus, understanding and modeling the relationship between forecast correction and combination is of importance and can result in guidelines when to use which method and when to apply both.

Lastly, forecast correction as well as combination only increase accuracy on average. Although the accuracy is often substantially increased, a decrease of the accuracy must be expected in a substantial number of cases. Substantial decreases must especially be expected directly after structural changes, when the forecast correction or combination model does not yet have enough new data to reasonably address the change, or in cases where the expert has knowledge for instance about special events. Although the robustness of the applied methods can be increased to avoid decreasing forecast accuracy as much as possible, decreases of the accuracy cannot be eliminated completely and are an inevitable result of the integration with statistical models. This aspect can only be addressed by checking with the expert who produced the judgmental forecast in cases where the difference to the original forecast is large. Currently, a project is underway with the sample company of the case study, in which a novel forecast support system is conceptualized and implemented that enables this interaction.

In summary, various directions for future research on the integration of judgmental forecasts with statistical models exist, which promise to provide new insights and to further improve performance in practice. This thesis provided a valuable basis for the theoretical understanding and improvement of forecast correction and combination, which enables promising future research as well as application in practice.

Part V

Appendix

Appendix A

Proofs

Proof of Theorem 2.1. The decomposition of the MSE can be derived as shown in Equation A.1.

$$\begin{aligned}
MSE =& \mathrm{E}\left[(A-F)^2\right] \\
&= (\mathrm{E}[A-F])^2 + \mathrm{Var}[A-F] \\
&= (\mathrm{E}[A] - \mathrm{E}[F])^2 + \mathrm{Var}[A] + \mathrm{Var}[F] - 2\mathrm{Cov}(F,A) \\
&= (\mu_A - \mu_F)^2 + \sigma_A^2 + \sigma_F^2 - 2\rho\sigma_A\sigma_F \\
&= (\mu_A - \mu_F)^2 + (\sigma_F - \rho\sigma_A)^2 + \sigma_A r - \rho^2\sigma_A^2 \\
&= (\mu_A - \mu_F)^2 + (\sigma_F - \rho\sigma_A)^2 + \left(1-\rho^2\right)\sigma_A^2 \qquad \text{(A.1)}
\end{aligned}$$

□

Proof of Theorem 2.2. In an ordinary least squares regression, the coefficients are defined as

$$\beta_1 = \rho\frac{\sigma_A}{\sigma_F}$$

and

$$\beta_0 = \mu_A - \beta_1\mu_F$$

Calculating the MSE of the derived corrected forecast, as shown in Equation A.2, shows that only the random (unsystematic) component of the MSE remains.

$$\begin{aligned}
MSE =& \mathrm{E}\left[(A - (\beta_0 + \beta_1 F))^2\right] \\
=& \mathrm{E}\left[((A-\mu_A) - \beta_1(F-\mu_F))^2\right] \\
=& \mathrm{E}\left[(A-\mu_A)^2\right] - 2\beta_1\mathrm{E}\left[(A-\mu_A)(F-\mu_F)\right] + \beta_1^2\mathrm{E}\left[(F-\mu_F)^2\right]
\end{aligned}$$

$$\begin{aligned}
&=\sigma_A^2 - 2\rho\frac{\sigma_A}{\sigma_F}\rho\sigma_A\sigma_F + \rho^2\frac{\sigma_A^2}{\sigma_F^2}\sigma_F^2\\
&=\sigma_A^2 - 2\rho^2\sigma_A^2 + \rho^2\sigma_A^2\\
&=(1-\rho^2)\sigma_A^2 \qquad \text{(A.2)}
\end{aligned}$$

□

Proof of Theorem 3.1. First, the expectation can easily be found as display in Equation A.3.

$$\begin{aligned}
&\mathrm{E}\left[A - F_C\right]\\
=&\mathrm{E}\left[(A-\hat{\mu}_A) - \hat{\beta}_1\left(F-\hat{\mu}_F\right)\right]\\
=&\mathrm{E}\left[A\right] - \mathrm{E}\left[\hat{\mu}_A\right] - \mathrm{E}\left[\hat{\beta}_1\right]\mathrm{E}\left[F-\hat{\mu}_F\right] - \underbrace{\mathrm{Cov}\left(\hat{\beta}_1, F-\hat{\mu}_F\right)}_{0}\\
=&\mu_A - \mu_A - \beta_1\mu_F + \beta_1\mu_F\\
=&0 \qquad \text{(A.3)}
\end{aligned}$$

Second, the variance of the error can be derived as presented in Equation A.4.

$$\begin{aligned}
&\mathrm{Var}\left[(A-\hat{\mu}_A) - \hat{\beta}_1\left(F-\hat{\mu}_F\right)\right]\\
=&\mathrm{Var}\left[A-\hat{\mu}_A\right] + \mathrm{Var}\left[\hat{\beta}_1\left(F-\hat{\mu}_F\right)\right] + \mathrm{Cov}\left(A-\hat{\mu}_A, \hat{\beta}_1\left(F-\hat{\mu}_F\right)\right)\\
=&\sigma_A^2 + \mathrm{Var}\left[\hat{\beta}_1\right]\mathrm{Var}\left[F-\hat{\mu}_F\right] + \mathrm{Var}\left[\hat{\beta}_1\right]\underbrace{\left(\mathrm{E}\left[F-\hat{\mu}_F\right]\right)^2}_{0}\\
&+\left(\mathrm{E}\left[\hat{\beta}_1\right]\right)^2\mathrm{Var}\left[F-\hat{\mu}_F\right] + \mathrm{E}\left[\hat{\beta}_1\right]\mathrm{Cov}\left(A-\hat{\mu}_A, F-\hat{\mu}_F\right)\\
=&\sigma_A^2 + \mathrm{Var}\left[\hat{\beta}_1\right]\sigma_F^2 + \left(\mathrm{E}\left[\hat{\beta}_1\right]\right)^2\sigma_F^2 + \mathrm{E}\left[\hat{\beta}_1\right]\mathrm{Cov}\left(A,F\right)\\
=&\sigma_A^2 + \sigma_F^2\left(\mathrm{Var}\left[\hat{\beta}_1\right] + \left(\mathrm{E}\left[\hat{\beta}_1\right]\right)^2\right) - 2\mathrm{E}\left[\hat{\beta}_1\right]\rho\sigma_A\sigma_F \qquad \text{(A.4)}
\end{aligned}$$

As the MSE can be decomposed into the squared expected error plus the error variance, the expected MSE is equal to the error variance. □

Proof of Theorem 3.2. First, $\mathrm{E}\left[\hat{\beta}_1\right] = \rho\frac{\sigma_A}{\sigma_F}$ since the OLS estimator is the best linear unbiased estimator (BLUE). Regarding the variance of the estimate, Pearson (1926) and Romanovsky (1926) independently found that $\mathrm{Var}\left[\hat{\beta}_1\right] = \frac{\sigma_A^2}{\sigma_F^2}\frac{1-\rho^2}{n-3}$ in case of multivariate normality.

Plugging expectation and variance of the coefficient estimate into the expected MSE in Theorem 3.1 yields the expected MSE of the corrected forecast as shown in Equation A.5.

$$\begin{aligned}\operatorname{Var}\left[A-F_C\right] =&\sigma_A^2+\sigma_F^2\left(\frac{\sigma_A^2}{\sigma_F^2}\frac{1-\rho^2}{n-3}+\rho^2\frac{\sigma_A^2}{\sigma_F^2}\right)-2\rho\frac{\sigma_A}{\sigma_F}\rho\sigma_A\sigma_F\\ =&\sigma_A^2+\frac{1-\rho^2}{n-3}\sigma_A^2+\rho^2\sigma_A^2-2\rho^2\sigma_A^2\\ =&\sigma_A^2+\frac{1-\rho^2}{n-3}\sigma_A^2-\rho^2\sigma_A^2\\ =&\left(1+\frac{1}{n-3}\right)\left(1-\rho^2\right)\sigma_A^2 \end{aligned}\tag{A.5}$$

□

Proof of Theorem 3.3. Reconsidering that the MSE of the original forecast is, in its decomposed form, $\left(\mu_A-\mu_F\right)^2+\left(\sigma_F-\rho\sigma_A\right)^2+\left(1-\rho^2\right)\sigma_A^2$. Equalizing the MSE with the result of Theorem 3.2 and solving for $\mathring{n}$ yields the minimal training sample size as shown in Equation A.6.

$$\begin{aligned}\left(1+\frac{1}{\mathring{n}-3}\right)\left(1-\rho^2\right)\sigma_A^2 =&\left(\mu_A-\mu_F\right)^2+\left(\sigma_F-\rho\sigma_A\right)^2+\left(1-\rho^2\right)\sigma_A^2\\ \frac{1}{\mathring{n}-3}\left(1-\rho^2\right)\sigma_A^2 =&\left(\mu_A-\mu_F\right)^2+\left(\sigma_F-\rho\sigma_A\right)^2\\ \frac{1}{\mathring{n}-3} =&\frac{\left(\mu_A-\mu_F\right)^2+\sigma_F^2-2\rho\sigma_A\sigma_F+\rho^2\sigma_A^2}{\left(1-\rho^2\right)\sigma_A^2}\\ \mathring{n} =&\frac{\left(1-\rho^2\right)\sigma_A^2}{\left(\mu_A-\mu_F\right)^2+\left(\sigma_F-\rho\sigma_A\right)^2}+3 \end{aligned}\tag{A.6}$$

The theorem then results by rounding the result up to the next integer since fractional or real-valued sample sizes do not exist. □

Proof of Theorem 3.4. Plugging $\mu_A=\tilde{\mu}_A$, $\mu_F=\tilde{\mu}_F-\Delta\mu$, and $\beta_1=\tilde{\beta}_1$ into Equation 3.5 yields the theorem as shown in Equation A.7.

$$\begin{aligned}MSE =&\mathrm{E}\left[\left(\left(\tilde{A}-\tilde{\mu}_A\right)-\tilde{\beta}_1\left(\tilde{F}-\tilde{\mu}_F-\Delta_\mu\right)\right)^2\right]\\ =&\mathrm{E}\left[\left(\left(\tilde{A}-\tilde{\mu}_A\right)-\tilde{\beta}_1\left(\tilde{F}-\tilde{\mu}_F\right)+\tilde{\beta}_1\Delta_\mu\right)^2\right]\\ =&\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2+2\tilde{\beta}_1\Delta_\mu\mathrm{E}\left[\tilde{A}-\tilde{\mu}_A\right]-2\tilde{\beta}_1^2\Delta_\mu\mathrm{E}\left[\tilde{F}-\tilde{\mu}_F\right]+\left(\tilde{\beta}_1\Delta_\mu\right)^2\\ =&\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2+\left(\tilde{\beta}_1\Delta_\mu\right)^2 \end{aligned}$$

$$= \left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 + \left(\tilde{\rho} \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F} \Delta_\mu\right)^2 \tag{A.7}$$

□

Proof of Theorem 3.5. Comparing the MSE for a correlation change in Theorem 3.4 with the decomposed MSE of the original forecast and solving for $\mathring{\Delta}_\mu$ yields in the critical value as shown in Equation A.8.

$$\begin{aligned}
\left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 + \left(\tilde{\beta}_1 \mathring{\Delta}_\mu\right)^2 &= \left(\tilde{\mu}_A - \tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A\right)^2 + \left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 \\
\left(\tilde{\beta}_1 \mathring{\Delta}_\mu\right)^2 &= \left(\tilde{\mu}_A - \tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A\right)^2 \\
\mathring{\Delta}_\mu &= \pm \sqrt{\frac{\left(\tilde{\mu}_A - \tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A\right)^2}{\tilde{\beta}_1^2}} \\
\mathring{\Delta}_\mu &= \pm \frac{\tilde{\sigma}_F}{\tilde{\rho}\tilde{\sigma}_A} \sqrt{\left(\tilde{\mu}_A - \tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F - \tilde{\rho}\tilde{\sigma}_A\right)^2}
\end{aligned} \tag{A.8}$$

□

Proof of Theorem 3.6. Plugging $\mu_A = \tilde{\mu}_A$, $\mu_F = \tilde{\mu}_F$, and $\beta_1 = \rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F} = \left(\tilde{\rho} - \Delta_\rho\right) \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F} = \tilde{\beta}_1 - \Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F}$ into Equation 3.5 yields the impact as shown in Equation A.9.

$$\begin{aligned}
MSE =& \mathrm{E}\left[\left(\left(\tilde{A} - \tilde{\mu}_A\right) - \left(\tilde{\beta}_1 - \Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F}\right)\left(\tilde{F} - \tilde{\mu}_F\right)\right)^2\right] \\
=& \mathrm{E}\left[\left(\left(\tilde{A} - \tilde{\mu}_A\right) - \tilde{\beta}_1\left(\tilde{F} - \tilde{\mu}_F\right) + \Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F}\left(\tilde{F} - \tilde{\mu}_F\right)\right)^2\right] \\
=& \left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 + 2\Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F} \mathrm{E}\left[\left(\tilde{A} - \tilde{\mu}_A\right)\left(\tilde{F} - \tilde{\mu}_F\right)\right] \\
& - 2\tilde{\beta}_1 \Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F} \mathrm{E}\left[\left(\tilde{F} - \tilde{\mu}_F\right)^2\right] + \left(\Delta_\rho \frac{\tilde{\sigma}_A}{\tilde{\sigma}_F}\right)^2 \mathrm{E}\left[\left(\tilde{F} - \tilde{\mu}_F\right)^2\right] \\
=& \left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 + 2\tilde{\beta}_1 \Delta_\rho \tilde{\sigma}_A \tilde{\sigma}_F - 2\tilde{\beta}_1 \Delta_\rho \tilde{\sigma}_A \tilde{\sigma}_F + \left(\Delta_\rho \tilde{\sigma}_A\right)^2 \\
=& \left(1 - \tilde{\rho}^2\right) \tilde{\sigma}_A^2 + \left(\Delta_\rho\right)^2 \tilde{\sigma}_A^2
\end{aligned} \tag{A.9}$$

□

Proof of Theorem 3.7. Equalizing the MSE for a correlation change in Theorem 3.6 with the decomposed MSE of the original forecast and solving for $\mathring{\Delta}_\rho$ results in the critical value derived in Equation A.10.

$$\begin{aligned}\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2+\left(\mathring{\Delta}_\rho\right)^2\tilde{\sigma}_A^2 &= \left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2+\left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2+\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2\\ \left(\mathring{\Delta}_\rho\right)^2\tilde{\sigma}_A^2 &= \left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2+\left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2\\ \mathring{\Delta}_\rho &= \pm\sqrt{\frac{\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2+\left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2}{\tilde{\sigma}_A^2}}\\ \mathring{\Delta}_\rho &= \pm\frac{1}{\tilde{\sigma}_A}\sqrt{\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2+\left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2} \qquad \text{(A.10)}\end{aligned}$$

□

Proof of Theorem 3.8. First note the relationship between β_1 and $\tilde{\beta}_1$ derived in Equation A.11.

$$\beta_1=\rho\frac{\sigma_A}{\sigma_F}=\tilde{\rho}\frac{\tilde{\sigma}_A}{\tilde{\sigma}_F-\Delta_\sigma}=\tilde{\beta}_1\frac{\tilde{\sigma}_F}{\tilde{\sigma}_F-\Delta_\sigma}=\tilde{\beta}_1+\tilde{\beta}_1\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma} \qquad \text{(A.11)}$$

Plugging this relationship, $\mu_A=\tilde{\mu}_A$, and $\mu_F=\tilde{\mu}_F$ into Equation 3.5 yields the theorem as shown in Equation A.12.

$$\begin{aligned}MSE =&\mathrm{E}\left[\left(\left(\tilde{A}-\tilde{\mu}_A\right)-\left(\tilde{\beta}_1+\tilde{\beta}_1\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\right)\left(\tilde{F}-\tilde{\mu}_F\right)\right)^2\right]\\ =&\mathrm{E}\left[\left(\left(\tilde{A}-\tilde{\mu}_A\right)-\tilde{\beta}_1\left(\tilde{F}-\tilde{\mu}_F\right)-\tilde{\beta}_1\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\left(\tilde{F}-\tilde{\mu}_F\right)\right)^2\right]\\ =&\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2-2\tilde{\beta}_1\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\mathrm{E}\left[\left(\tilde{A}-\tilde{\mu}_A\right)\left(\tilde{F}-\tilde{\mu}_F\right)\right]\\ &+2\tilde{\beta}_1^2\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\mathrm{E}\left[\left(\tilde{F}-\tilde{\mu}_F\right)^2\right]+\left(\tilde{\beta}_1\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\right)^2\mathrm{E}\left[\left(\tilde{F}-\tilde{\mu}_F\right)^2\right]\\ =&\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2-2\tilde{\beta}_1^2\tilde{\sigma}_F^2\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\\ &+2\tilde{\beta}_1^2\tilde{\sigma}_F^2\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}+\left(\tilde{\rho}\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\right)^2\left(\tilde{\sigma}_A\right)^2\\ =&\left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2+\left(\tilde{\rho}\frac{\Delta_\sigma}{\tilde{\sigma}_F-\Delta_\sigma}\right)^2\tilde{\sigma}_A^2 \qquad \text{(A.12)}\end{aligned}$$

□

Proof of Theorem 3.9. Equalizing the MSE under a forecast variance change (Theorem 3.8) with the decomposed MSE of the original forecast yields the critical change as shown in Equation A.13.

$$
\begin{aligned}
0 &= \left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2 + \left(\tilde{\rho}\frac{\mathring{\Delta}_\sigma}{\tilde{\sigma}_F-\mathring{\Delta}_\sigma}\right)^2 \tilde{\sigma}_A^2 - \left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 - \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2 - \left(1-\tilde{\rho}^2\right)\tilde{\sigma}_A^2\\
0 &= \left(\tilde{\rho}\frac{\mathring{\Delta}_\sigma}{\tilde{\sigma}_F-\mathring{\Delta}_\sigma}\right)^2 \tilde{\sigma}_A^2 - \left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 - \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2\\
0 &= \left(\mathring{\Delta}_\sigma\right)^2\left(\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2 - \tilde{\rho}^2\tilde{\sigma}_A^2\right)\\
&\quad - \mathring{\Delta}_\sigma\left(2\tilde{\sigma}_F\left(\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2\right)\right)\\
&\quad + \tilde{\sigma}_F^2\left(\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2\right)\\
\mathring{\Delta}_\sigma &= \frac{\sigma_F\left(\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2\right) \pm \tilde{\rho}\tilde{\sigma}_A\sqrt{\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2}}{\left(\tilde{\mu}_A-\tilde{\mu}_F\right)^2 + \left(\tilde{\sigma}_F-\tilde{\rho}\tilde{\sigma}_A\right)^2 - \tilde{\rho}^2\tilde{\sigma}_A^2} \qquad \text{(A.13)}
\end{aligned}
$$

□

Proof of Theorem 3.10. The variances of the modified forecasts and actuals with added systematic component T are displayed in Equations A.14 and A.15.

$$\sigma_{A'}^2 = \sigma_A^2 + \sigma_T^2 \qquad \text{(A.14)}$$

$$\sigma_{F'}^2 = \sigma_F^2 + \sigma_F^2 \qquad \text{(A.15)}$$

The covariance between the modified forecasts and actuals is

$$
\begin{aligned}
\operatorname{Cov}\left(A',F'\right) &= \operatorname{Cov}\left(A+T,F+T\right)\\
&= \operatorname{Cov}\left(A,F\right) + \underbrace{\operatorname{Cov}\left(A,T\right)}_{0} + \underbrace{\operatorname{Cov}\left(F,T\right)}_{0} + \sigma_T^2\\
&= \rho\sigma_A\sigma_F + \sigma_T^2 \qquad \text{(A.16)}
\end{aligned}
$$

Using Equations A.14, A.15, A.16, the regression coefficient β_1' for the actuals and forecasts with the systematic component can be calculated as shown in Equation A.17.

$$\beta_1' = \frac{\operatorname{Cov}\left(A',F'\right)}{\sigma_{F'}^2} = \frac{\rho\sigma_A\sigma_F + \sigma_T^2}{\sigma_F^2+\sigma_T^2} \qquad \text{(A.17)}$$

Clearly, $\lim_{\sigma_T\to\infty}\beta_1' = 1$, indicating that regression bias vanishes for increasing variance of the systematic component. □

Proof of Theorem 4.1. The expectation of shrinked weights is a linear combination of the expectations of SA and ow, as shown in Equation A.18.

$$\begin{aligned}\mathrm{E}\left[\hat{w}^{\lambda}\right] &= \mathrm{E}\left[\lambda w^{S} + (1-\lambda)\,\hat{w}^{O}\right] \\ &= \lambda \mathrm{E}\left[w^{S}\right] + (1-\lambda)\,\mathrm{E}\left[\hat{w}^{O}\right] \qquad \text{(A.18)}\end{aligned}$$

Since the weights are fixed for SA and thus not a random variable, the expectation $\mathrm{E}\left[w^{S}\right]$ is directly determined by k. For the estimation of OW, Granger and Ramanathan (1984) proved an equivalence to a multiple ordinary least squares (OLS) linear regression. Since the OLS linear regression estimator is the best linear unbiased estimator (BLUE), the optimal weights estimator must necessarily be unbiased. As a consequence, simply plugging the definitions of SA and OW into Equation A.18 yields Equation A.19.

$$\mathrm{E}\left[\hat{w}^{\lambda}\right] = \lambda\frac{1}{k}\vec{1} + (1-\lambda)\,\frac{\Sigma^{-1}\vec{1}}{\vec{1}^{\top}\Sigma^{-1}\vec{1}} \qquad \text{(A.19)}$$

□

Proof of Theorem 4.2. With $w_1^O = \mathrm{E}\left[\hat{w}_1^O\right] = \frac{\sigma_2^2-\rho\sigma_1\sigma_2}{\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2}$ (Bates and Granger, 1969), the expectation of the shrinked weights in the bivariate case is introduced in Equation A.20.

$$\begin{aligned}\mathrm{E}\left[\hat{w}^{\lambda}\right] &= \left(\frac{\lambda}{2} + (1-\lambda)\,\mathrm{E}\left[\hat{w}_1^O\right], \frac{\lambda}{2} + (1-\lambda)\,\mathrm{E}\left[\hat{w}_2^O\right]\right) \\ &= \left(\frac{\lambda}{2} + (1-\lambda)\,w_1^O, \frac{\lambda}{2} + (1-\lambda)\,w_2^O\right) \\ &= \left(\frac{\lambda}{2} + \frac{(1-\lambda)\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)}{\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2}, \frac{\lambda}{2} + \frac{(1-\lambda)\left(\sigma_1^2-\rho\sigma_1\sigma_2\right)}{\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2}\right) \\ &= \left(\frac{\lambda\sigma_1^2 + (2-\lambda)\,\sigma_2^2 - 2\rho\sigma_1\sigma_2}{2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}, \frac{(2-\lambda)\,\sigma_1^2 + \lambda\sigma_2^2 - 2\rho\sigma_1\sigma_2}{2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}\right) \\ &= \left(\frac{\frac{\lambda}{2}\left(\sigma_1^2-\sigma_2^2\right) + \sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}, \frac{\frac{\lambda}{2}\left(\sigma_2^2-\sigma_1^2\right) + \sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\right) \qquad \text{(A.20)}\end{aligned}$$

□

Proof of Theorem 4.3. If Ω, the sampling covariance matrix of OW, is known, the sampling covariance matrix $\Omega^{\lambda} \in \mathbb{R}^{k\times k}$ of the elements in $\hat{w}^{\lambda}$ can easily be derived, as shown in Equation A.21.

$$\begin{aligned}
\Omega^{\lambda} =& \mathrm{Cov}\left(\hat{w}^{\lambda}, \hat{w}^{\lambda}\right) \\
=& \mathrm{Cov}\left(\lambda w^{S} + (1-\lambda)\,\hat{w}^{O}, \lambda w^{S} + (1-\lambda)\,\hat{w}^{O}\right) \\
=& \lambda^{2} \underbrace{\mathrm{Cov}\left(w^{S}, w^{S}\right)}_{0} + 2\lambda\,(1-\lambda) \underbrace{\mathrm{Cov}\left(w^{S}, \hat{w}^{O}\right)}_{0} + (1-\lambda)^{2}\,\mathrm{Cov}\left(\hat{w}^{O}, \hat{w}^{O}\right) \\
=& (1-\lambda)^{2}\,\mathrm{Cov}\left(\hat{w}^{O}, \hat{w}^{O}\right) \\
=& (1-\lambda)^{2}\,\Omega
\end{aligned} \tag{A.21}$$

Hence, the challenge reduces to deriving the sampling covariance matrix of $\hat{w}^{O}$, which can be done by utilizing the equivalence of OW and the coefficients of a multiple linear regression shown by Granger and Ramanathan (1984). They have proven that the weight estimates $\hat{w}^{O}_{1}, \ldots, \hat{w}^{O}_{k-1}$ can be calculated by linearly regressing the error of the k-th forecast, E_k, on the differences to the other errors, as shown in Equation A.22. Although an intercept term is included when calculating the coefficients, the intercept can then be omitted because of the unbiasedness-assumption of the individual forecasts and the resulting true intercept of zero.

$$E_k = w^{O}_{1}\,(E_k - E_1) + \cdots + w^{O}_{k-1}\,(E_k - E_{k-1}) + \epsilon \tag{A.22}$$

The covariance matrix of independent and the dependent variables can be found to be the defined modified covariance matrix Σ'. More precisely, Σ'_{11} is the covariance between the independent variables, $\Sigma_{k,k}$ the variance of the dependent variable (the error of the k-th forecast) and Σ'_{12} the covariance between the dependent variable and the independent variables.

Kshirsagar (1961) derived the distribution of the coefficients of a multiple regression by specifying the density. The density corresponds to a matrix t-distribution. Consequently, because of the equivalence between optimal weights combination and linear regression, the sampling distribution of $\left(\hat{w}^{O}_{1}, \ldots, \hat{w}^{O}_{k-1}\right)$ is a matrix t-distribution with $n-k+1$ degrees of freedom, and scales $(\Sigma'_{11})^{-1}$ and $\Sigma_{k,k} - (\Sigma'_{12})^{\top}(\Sigma'_{11})^{-1}\Sigma'_{12}$.

The sampling covariance matrix of the first $k-1$ weights follows from the results of Gupta and Nagar (1999) on the properties of matrix variate t-distributions and is shown in Equation A.23.

$$\Omega' = \frac{1}{n-k-1}\left(\Sigma'_{11}\right)^{-1}\left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top}\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}\right) \tag{A.23}$$

Since this sampling covariance matrix only includes the first $k-1$, the matrix has to be further modified to include the last forecast. Trivially, the estimate for the last weight $\hat{w}_k^O$ suffices Equation A.24.

$$\hat{w}_k^O = 1 - \sum_{i=1}^{k-1} \hat{w}_i^O \tag{A.24}$$

Using this relationship, the complete sampling covariance matrix can easily be derived. Since $\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top}\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}$ is a scalar value, adding the last forecast reduces to extending $\left(\Sigma'_{11}\right)^{-1}$ appropriately. The required augmented matrix is clearly the defined matrix Ω^O. Replacing the matrix yields the sampling covariance matrix of optimal weights in Equation A.25.

$$\Omega = \frac{1}{n-k-1}\Omega^O\left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top}\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}\right) \tag{A.25}$$

Plugging Ω into Equation A.21 proves the claimed sampling covariance matrix. □

Proof of Theorem 4.4. In a first step, the bivariate case of the modified covariance matrix Σ', as presented in Equation A.26, can be directly found by simply plugging in the definitions of the elements of Σ.

$$\Sigma' = \begin{bmatrix} \sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2 & \sigma_2^2 - \rho\sigma_1\sigma_2 \\ \sigma_2^2 - \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{A.26}$$

Using the general definition of Ω^O and the bivariate variant of Σ', Ω^O in the bivariate case is defined as shown in Equation A.27.

$$\Omega^O = \frac{1}{\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{A.27}$$

The weight covariance matrix can then be derived as shown in Equation A.28.

$$\Omega^\lambda = \frac{(1-\lambda)^2}{n-k-1}\Omega^O\left(\Sigma_{k,k} - \left(\Sigma'_{12}\right)^{\top}\left(\Sigma'_{11}\right)^{-1}\Sigma_{12}\right)$$

$$
\begin{aligned}
&= \frac{(1-\lambda)^2}{n-3} \Omega^O \left(\sigma_2^2 - \frac{\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)^2}{\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2} \right) \\
&= \frac{(1-\lambda)^2}{n-3} \Omega^O \frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2} \\
&= \frac{(1-\lambda)^2}{n-3} \frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_2^2 - 2\rho\sigma_1\sigma_2 + \sigma_1^2\right)^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad \text{(A.28)}
\end{aligned}
$$

□

Proof of Theorem 4.5. Using the result on the covariance of products of random variables with vanishing third moments by Bohrnstedt and Goldberger (1969), the combined error variance can be derived as presented in Equation A.29.

$$
\begin{aligned}
&\operatorname{Var}\left[\left(\hat{w}^\lambda\right)^\top \tilde{E}\right] = \sum_{i,j} \operatorname{Cov}\left(\hat{w}_i^\lambda \tilde{E}_i, \hat{w}_j^\lambda \tilde{E}_j\right) \\
&= \sum_{i,j} \underbrace{\mathrm{E}\left[\tilde{E}_i\right] \mathrm{E}\left[\tilde{E}_j\right] \operatorname{Cov}\left(\hat{w}_i^\lambda, \hat{w}_j^\lambda\right)}_{0} + \sum_{i,j} \underbrace{\mathrm{E}\left[\tilde{E}_i\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \operatorname{Cov}\left(\hat{w}_i^\lambda, \tilde{E}_j\right)}_{0} \\
&\quad + \sum_{i,j} \underbrace{\mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\tilde{E}_j\right] \operatorname{Cov}\left(\tilde{E}_i, \hat{w}_j^\lambda\right)}_{0} + \sum_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \operatorname{Cov}\left(\tilde{E}_i, \tilde{E}_j\right) \\
&\quad + \sum_{i,j} \operatorname{Cov}\left(\tilde{E}_i, \tilde{E}_j\right) \operatorname{Cov}\left(\hat{w}_i^\lambda, \hat{w}_j^\lambda\right) + \sum_{i,j} \underbrace{\operatorname{Cov}\left(\tilde{E}_i, \hat{w}_j^\lambda\right) \operatorname{Cov}\left(\hat{w}_i^\lambda, \tilde{E}_j\right)}_{0 \text{ (by assumption)}} \\
&= \sum_{i,j} \operatorname{Cov}\left(\tilde{E}_i, \tilde{E}_j\right) \operatorname{Cov}\left(\hat{w}_i^\lambda, \hat{w}_j^\lambda\right) + \sum_{i,j} \operatorname{Cov}\left(\tilde{E}_i, \tilde{E}_j\right) \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&= \sum_{i,j} \tilde{\Sigma}_{i,j} \Omega_{i,j}^\lambda + \sum_{i,j} \tilde{\Sigma}_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \qquad \text{(A.29)}
\end{aligned}
$$

□

Proof of Theorem 4.6. Plugging the expectation and sampling covariance of shrinked weights in the bivariate case introduced in Theorems 4.2 and 4.4 into the result for the general case (Theorem 4.5) yields the combined error variance as derived in Equation A.30.

$$
\operatorname{Var}\left[\left(\hat{w}^\lambda\right)^\top \tilde{E}\right] = \sum_{i,j} \tilde{\Sigma}_{i,j} \Omega_{i,j}^\lambda + \sum_{i,j} \tilde{\Sigma}_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right]
$$

$$
\begin{aligned}
=&\frac{(1-\lambda)^2}{n-3}\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2\right)^2}\left(\tilde{\sigma}_1^2+2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2+\tilde{\sigma}_2^2\right)\\
&+\tilde{\sigma}_1^2\left(\frac{\frac{\lambda}{2}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\right)^2\\
&+\tilde{\sigma}_2^2\left(\frac{\frac{\lambda}{2}\left(\sigma_2^2-\sigma_1^2\right)+\sigma_1^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\right)^2\\
&+2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2\frac{\frac{\lambda}{2}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\frac{\frac{\lambda}{2}\left(\sigma_2^2-\sigma_1^2\right)+\sigma_1^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\\
=&\frac{1}{\left(\sigma_2^2-2\rho\sigma_1\sigma_2+\sigma_1^2\right)^2}\Bigg(\\
&\frac{(1-\lambda)^2}{n-3}\left(1-\rho^2\right)\sigma_1^2\sigma_2^2\left(\tilde{\sigma}_1^2+2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2+\tilde{\sigma}_2^2\right)\\
&+\tilde{\sigma}_1^2\left(\frac{\lambda}{2}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2\right)^2\\
&+\tilde{\sigma}_2^2\left(\frac{\lambda}{2}\left(\sigma_2^2-\sigma_1^2\right)+\sigma_1^2-\rho\sigma_1\sigma_2\right)^2\\
&+2\tilde{\rho}\tilde{\sigma}_1\tilde{\sigma}_2\left(\frac{\lambda}{2}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2\right)\left(\frac{\lambda}{2}\left(\sigma_2^2-\sigma_1^2\right)+\sigma_1^2-\rho\sigma_1\sigma_2\right)\Bigg)
\end{aligned}
\tag{A.30}
$$

□

Proof of Theorem 4.7. Equalizing the expected combined out-of-sample error variances of two combinations with λ_1 and λ_2 and solving for $\mathring{n}$ yields the minimal sample size, as presented in Equation A.31.

$$
\begin{aligned}
0 =&\operatorname{Var}\left[\left(\hat{w}^{\lambda_1}\right)^\top\tilde{E}\right]-\operatorname{Var}\left[\left(\hat{w}^{\lambda_2}\right)^\top\tilde{E}\right]\\
0 =&\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\Omega_{i,j}^{\lambda_1}-\Omega_{i,j}^{\lambda_2}\right)+\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\mathrm{E}\left[\hat{w}_i^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_i^{\lambda_2}\right]\right)\left(\mathrm{E}\left[\hat{w}_j^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_j^{\lambda_2}\right]\right)\\
0 =&\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\left(\frac{(1-\lambda_1)^2}{\mathring{n}-k-1}-\frac{(1-\lambda_2)^2}{\mathring{n}-k-1}\right)\Omega_{i,j}^{O}\left(\Sigma_{k,k}-\left(\Sigma'_{12}\right)^\top\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}\right)\right)\\
&+\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\mathrm{E}\left[\hat{w}_i^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_i^{\lambda_2}\right]\right)\left(\mathrm{E}\left[\hat{w}_j^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_j^{\lambda_2}\right]\right)\\
0 =&\left(\frac{(1-\lambda_1)^2}{\mathring{n}-k-1}-\frac{(1-\lambda_2)^2}{\mathring{n}-k-1}\right)\sum_{i,j}\tilde{\Sigma}_{i,j}\Omega_{i,j}^{O}\left(\Sigma_{k,k}-\left(\Sigma'_{12}\right)^\top\left(\Sigma'_{11}\right)^{-1}\Sigma'_{12}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\mathrm{E}\left[\hat{w}_i^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_i^{\lambda_2}\right]\right)\left(\mathrm{E}\left[\hat{w}_j^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_j^{\lambda_2}\right]\right)\\
\mathring{n} =&\frac{\left((1-\lambda_2)^2-(1-\lambda_1)^2\right)\sum_{i,j}\tilde{\Sigma}_{i,j}\Omega^O_{i,j}\left(\Sigma_{k,k}-\left(\Sigma'_{12}\right)^\top\left(\Sigma'_{11}\right)^{-1}\Sigma_{12}\right)}{\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\mathrm{E}\left[\hat{w}_i^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_i^{\lambda_2}\right]\right)\left(\mathrm{E}\left[\hat{w}_j^{\lambda_1}\right]-\mathrm{E}\left[\hat{w}_j^{\lambda_2}\right]\right)}+k+1
\end{aligned}
\tag{A.31}
$$

□

Proof of Theorem 4.8. Plugging the previously derived formuale for the bivariate case into Equation 4.5 and simplifying results in the critical value as presented in Equation A.32.

$$
\begin{aligned}
\mathring{n} =&\frac{(1-\lambda)^2\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\Sigma_{k,k}-\left(\Sigma'_{12}\right)^\top\left(\Sigma'_{11}\right)^{-1}\Sigma_{12}\right)\Omega^O_{i,j}}{\sum_{i,j}\tilde{\Sigma}_{i,j}\left(\frac{1}{k}-\mathrm{E}\left[\hat{w}_i^{\lambda}\right]\right)\left(\frac{1}{k}-\mathrm{E}\left[\hat{w}_j^{\lambda}\right]\right)}+k+1\\
=&\frac{(1-\lambda)^2\sum_{i,j}\Sigma_{i,j}\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}}{\sigma_1^2\left(\frac{1}{2}-w_1^O\right)^2+\sigma_2^2\left(\frac{1}{2}-\left(1-w_1^O\right)\right)^2+2\rho\sigma_1\sigma_2\left(\frac{1}{2}-w_1^O\right)\left(\frac{1}{2}-\left(1-w_1^O\right)\right)}\\
&+3\\
=&\frac{(1-\lambda)^2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}}{\sigma_1^2\left(\frac{1}{2}-w_1^O\right)^2+\sigma_2^2\left(\frac{1}{2}-w_1^O\right)^2-2\rho\sigma_1\sigma_2\left(\frac{1}{2}-w_1^O\right)^2}+3\\
=&\frac{(1-\lambda)^2\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)\left(\frac{\sigma_2^2-\sigma_1^2}{2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}\right)^2}+3\\
=&(1-\lambda)^2\left(1-\rho^2\right)\frac{4\sigma_1^2\sigma_2^2}{\left(\sigma_2^2-\sigma_1^2\right)^2}+3\\
=&(1-\lambda)^2\left(1-\rho^2\right)\left(\frac{2\sigma_1\sigma_2}{\sigma_2^2-\sigma_1^2}\right)^2+3
\end{aligned}
\tag{A.32}
$$

□

Proof of Theorem 4.9. If the error correlation between forecasts p and q changes, only the elements p,q and q,p differ between Σ and $\tilde{\Sigma}$. The combined error variance can consequently be expressed using Σ while adjusting for the two changed elements.

Let σ_p^2, σ_q^2 denote the error variance of the two forecasts and $\rho_{p,q}$ the error correlation. Then the first term of the combined error variance in Theorem 4.5 can be reformulated as shown in Equation A.33.

$$\begin{aligned}
\sum_{i,j} \tilde{\Sigma}_{i,j}\Omega^{\lambda}_{i,j} &= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\Omega^{\lambda}_{i,j} + \tilde{\Sigma}_{p,q}\Omega^{\lambda}_{p,q} + \tilde{\Sigma}_{q,p}\Omega^{\lambda}_{q,p} \\
&= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\Omega^{\lambda}_{i,j} + 2\tilde{\Sigma}_{p,q}\Omega^{\lambda}_{p,q} \\
&= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\Omega^{\lambda}_{i,j} + 2\left(\Sigma_{p,q} + \Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\right)\Omega^{\lambda}_{p,q} \\
&= \sum_{i,j} \Sigma_{i,j}\Omega^{\lambda}_{i,j} + 2\Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\Omega^{\lambda}_{p,q}
\end{aligned} \tag{A.33}$$

Likewise, the second term in Theorem 4.5 can be reformulated as presented in Equation A.34.

$$\begin{aligned}
\sum_{i,j} \tilde{\Sigma}_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] &= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] + \tilde{\Sigma}_{p,q}\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right] \\
&\quad + \tilde{\Sigma}_{q,p}\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right] \\
&= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] + 2\tilde{\Sigma}_{p,q}\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right] \\
&= \sum_{i,j \notin \{p,q\}} \Sigma_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] \\
&\quad + 2\left(\Sigma_{p,q} + \Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\right)\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right] \\
&= \sum_{i,j} \Sigma_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] + 2\Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right]
\end{aligned} \tag{A.34}$$

Combining Equations A.33 and A.34 yields the claimed formulation as shown in Equation A.35.

$$\begin{aligned}
\mathrm{Var}\left[\left(\hat{w}^{\lambda}\right)^{\top}\tilde{E}\right] &= \sum_{i,j} \Sigma_{i,j}\Omega^{\lambda}_{i,j} + \sum_{i,j} \Sigma_{i,j}\mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right] \\
&\quad + 2\Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\Omega^{\lambda}_{p,q} + 2\Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\mathrm{E}\left[\hat{w}^{\lambda}_q\right] \\
&= \sum_{i,j} \Sigma_{i,j}\left(\Omega^{\lambda}_{i,j} + \mathrm{E}\left[\hat{w}^{\lambda}_i\right]\mathrm{E}\left[\hat{w}^{\lambda}_j\right]\right) + 2\Delta_{\rho}\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\left(\Omega^{\lambda}_{p,q} + \mathrm{E}\left[\hat{w}^{\lambda}_q\right]\right)
\end{aligned}$$

$$= \sum_{i,j} \Sigma_{i,j} \Psi^{\lambda}_{i,j} + 2\Delta_{\rho} \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Psi^{\lambda}_{p,q} \tag{A.35}$$

□

Proof of Theorem 4.10. Let σ_p^2 denote the error variance of the forecast. and $\rho_{p,q}$ the error correlation between forecast p and another forecast q. Then the first term of the combined error variance in Theorem 4.5 can be reformulated as shown in Equation A.36.

$$\begin{aligned}
\sum_{i,j} \tilde{\Sigma}_{i,j} \Omega^{\lambda}_{i,j} &= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + \sum_{j \neq p} \tilde{\Sigma}_{p,j} \Omega^{\lambda}_{p,j} + \sum_{i \neq p} \tilde{\Sigma}_{i,p} \Omega^{\lambda}_{i,p} + \tilde{\Sigma}_{p,p} \Omega^{\lambda}_{p,p} \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + 2 \sum_{j \neq p} \tilde{\Sigma}_{p,j} \Omega^{\lambda}_{p,j} + \tilde{\Sigma}_{p,p} \Omega^{\lambda}_{p,p} \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + 2 \sum_{j \neq p} \rho_{p,j} \tilde{\sigma}_p \sigma_j \Omega^{\lambda}_{p,j} + \tilde{\sigma}_p^2 \Omega^{\lambda}_{p,p} \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + 2 \sum_{j \neq p} \rho_{p,j} \left(\sigma_p + \Delta_{\sigma}\right) \sigma_j \Omega^{\lambda}_{p,j} + \left(\sigma_p + \Delta_{\sigma}\right)^2 \Omega^{\lambda}_{p,p} \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + 2 \sum_{j \neq p} \rho_{p,j} \sigma_p \sigma_j \Omega^{\lambda}_{p,j} + \sigma_p^2 \Omega^{\lambda}_{p,p} \\
&\quad + 2 \sum_{j \neq p} \rho_{p,j} \Delta_{\sigma} \sigma_j \Omega^{\lambda}_{p,j} + \left(2\sigma_p \Delta_{\sigma} + \Delta_{\sigma}^2\right) \Omega^{\lambda}_{p,p} \\
&= \sum_{i,j} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + 2 \sum_{j \neq p} \rho_{p,j} \Delta_{\sigma} \sigma_j \Omega^{\lambda}_{p,j} + \left(2\sigma_p \Delta_{\sigma} + \Delta_{\sigma}^2\right) \Omega^{\lambda}_{p,p} \\
&= \sum_{i,j} \Sigma_{i,j} \Omega^{\lambda}_{i,j} + \frac{2\Delta_{\sigma}}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Omega^{\lambda}_{p,j} + \left(2\sqrt{\Sigma_{p,p}} \Delta_{\sigma} + \Delta_{\sigma}^2\right) \Omega^{\lambda}_{p,p} \qquad \text{(A.36)}
\end{aligned}$$

Likewise, the second term in Theorem 4.5 can be reformulated as presented in Equation A.37.

$$\begin{aligned}
&\sum_{i,j} \tilde{\Sigma}_{i,j} \mathrm{E}\left[\hat{w}^{\lambda}_i\right] \mathrm{E}\left[\hat{w}^{\lambda}_j\right] \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \mathrm{E}\left[\hat{w}^{\lambda}_i\right] \mathrm{E}\left[\hat{w}^{\lambda}_j\right] + \sum_{j \neq p} \tilde{\Sigma}_{p,j} \mathrm{E}\left[\hat{w}^{\lambda}_p\right] \mathrm{E}\left[\hat{w}^{\lambda}_j\right] + \sum_{i \neq p} \tilde{\Sigma}_{i,p} \mathrm{E}\left[\hat{w}^{\lambda}_i\right] \mathrm{E}\left[\hat{w}^{\lambda}_p\right] \\
&\quad + \tilde{\Sigma}_{p,p} \left(\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\right)^2 \\
&= \sum_{i \neq p, j \neq p} \Sigma_{i,j} \mathrm{E}\left[\hat{w}^{\lambda}_i\right] \mathrm{E}\left[\hat{w}^{\lambda}_j\right] + 2 \sum_{j \neq p} \tilde{\Sigma}_{p,j} \mathrm{E}\left[\hat{w}^{\lambda}_p\right] \mathrm{E}\left[\hat{w}^{\lambda}_j\right] + \tilde{\Sigma}_{p,p} \left(\mathrm{E}\left[\hat{w}^{\lambda}_p\right]\right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\neq p, j\neq p} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + 2\sum_{j\neq p} \rho_{p,j}\tilde{\sigma}_p\sigma_j \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + \tilde{\sigma}_p^2 \left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&= \sum_{i\neq p, j\neq p} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + 2\sum_{j\neq p} \rho_{p,j}\left(\sigma_p + \Delta_\sigma\right)\sigma_j \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&\quad + \left(\sigma_p + \Delta_\sigma\right)^2 \left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&= \sum_{i\neq p, j\neq p} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + 2\sum_{j\neq p} \rho_{p,j}\sigma_p\sigma_j \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + \sigma_p^2 \left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&\quad + 2\sum_{j\neq p} \rho_{p,j}\Delta_\sigma\sigma_j \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + \left(2\sigma_p\Delta_\sigma + \Delta_\sigma^2\right)\left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&= \sum_{i,j} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + 2\sum_{j\neq p} \rho_{p,j}\Delta_\sigma\sigma_j \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&\quad + \left(2\sigma_p\Delta_\sigma + \Delta_\sigma^2\right)\left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&= \sum_{i,j} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j\neq p} \Sigma_{p,j} \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&\quad + \left(2\sqrt{\Sigma_{p,p}}\Delta_\sigma + \Delta_\sigma^2\right)\left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \qquad \text{(A.37)}
\end{aligned}
$$

Combining Equations A.36 and A.37 yields the claimed as shown in Equation A.38

$$
\begin{aligned}
\mathrm{Var}\left[\left(\hat{w}^\lambda\right)^\top \tilde{E}\right] &= \sum_{i,j} \Sigma_{i,j}\Omega_{i,j}^\lambda + \sum_{i,j} \Sigma_{i,j} \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&\quad + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j\neq p} \Sigma_{p,j}\Omega_{p,j}^\lambda + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j\neq p} \Sigma_{p,j} \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right] \\
&\quad + \left(2\sqrt{\Sigma_{p,p}}\Delta_\sigma + \Delta_\sigma^2\right)\Omega_{p,p}^\lambda + \left(2\sqrt{\Sigma_{p,p}}\Delta_\sigma + \Delta_\sigma^2\right)\left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2 \\
&= \sum_{i,j} \Sigma_{i,j}\left(\Omega_{i,j}^\lambda + \mathrm{E}\left[\hat{w}_i^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right]\right) \\
&\quad + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j\neq p} \Sigma_{p,j}\left(\Omega_{p,j}^\lambda + \mathrm{E}\left[\hat{w}_p^\lambda\right] \mathrm{E}\left[\hat{w}_j^\lambda\right]\right) \\
&\quad + \left(2\sqrt{\Sigma_{p,p}}\Delta_\sigma + \Delta_\sigma^2\right)\left(\Omega_{p,p}^\lambda + \left(\mathrm{E}\left[\hat{w}_p^\lambda\right]\right)^2\right) \\
&= \sum_{i,j} \Sigma_{i,j}\Psi_{i,j}^\lambda + \frac{2\Delta_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j\neq p} \Sigma_{p,j}\Psi_{p,j}^\lambda + \left(2\sqrt{\Sigma_{p,p}}\Delta_\sigma + \Delta_\sigma^2\right)\Psi_{p,p}^\lambda
\end{aligned}
$$

(A.38)

□

Proof of Theorem 4.11. Plugging λ_1 and λ_2 into Theorem 4.9 and equalizing results in the critical value as shown in Equation A.39.

$$\begin{aligned}
0 =& \mathrm{Var}\left[\left(\hat{w}^{\lambda_1}\right)^\top \tilde{E}\right] - \mathrm{Var}\left[\left(\hat{w}^{\lambda_2}\right)^\top \tilde{E}\right] \\
0 =& \sum_{i,j} \Sigma_{i,j} \Psi_{i,j}^{\lambda_1} + 2\mathring{\Delta}_\rho \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Psi_{p,q}^{\lambda_1} - \sum_{i,j} \Sigma_{i,j} \Psi_{i,j}^{\lambda_2} - 2\mathring{\Delta}_\rho \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Psi_{p,q}^{\lambda_2} \\
0 =& \sum_{i,j} \Sigma_{i,j} \left(\Psi_{i,j}^{\lambda_1} - \Psi_{i,j}^{\lambda_2}\right) + 2\mathring{\Delta}_\rho \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \left(\Psi_{p,q}^{\lambda_1} - \Psi_{p,q}^{\lambda_2}\right) \\
0 =& \sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1} + 2\mathring{\Delta}_\rho \sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Delta\Psi_{p,q}^{\lambda_1} \\
\mathring{\Delta}_\rho =& - \frac{\sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1,\lambda_2}}{2\sqrt{\Sigma_{p,p}\Sigma_{q,q}} \Delta\Psi_{p,q}^{\lambda_1,\lambda_2}}
\end{aligned} \tag{A.39}$$

□

Proof of Theorem 4.12. Plugging λ_1 and λ_2 into Theorem 4.10 and equalizing yields the critical value as displayed in Equation A.40.

$$\begin{aligned}
0 =& \mathrm{Var}\left[\left(\hat{w}^{\lambda_1}\right)^\top \tilde{E}\right] - \mathrm{Var}\left[\left(\hat{w}^{\lambda_2}\right)^\top \tilde{E}\right] \\
0 =& \sum_{i,j} \Sigma_{i,j} \Psi_{i,j}^{\lambda_1} - \sum_{i,j} \Sigma_{i,j} \Psi_{i,j}^{\lambda_2} + \frac{2\mathring{\Delta}_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Psi_{p,j}^{\lambda_1} - \frac{2\mathring{\Delta}_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Psi_{p,j}^{\lambda_2} \\
&+ \left(2\sqrt{\Sigma_{p,p}}\mathring{\Delta}_\sigma + \mathring{\Delta}_\sigma^2\right) \Psi_{p,p}^{\lambda_1} - \left(2\sqrt{\Sigma_{p,p}}\mathring{\Delta}_\sigma + \mathring{\Delta}_\sigma^2\right) \Psi_{p,p}^{\lambda_2} \\
0 =& \sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1,\lambda_2} + \frac{2\mathring{\Delta}_\sigma}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Delta\Psi_{p,j}^{\lambda_1,\lambda_2} + \left(2\sqrt{\Sigma_{p,p}}\mathring{\Delta}_\sigma + \mathring{\Delta}_\sigma^2\right) \Delta\Psi_{p,p}^{\lambda_1,\lambda_2} \\
0 =& \mathring{\Delta}_\sigma^2 \Delta\Psi_{p,p}^{\lambda_1,\lambda_2} + \mathring{\Delta}_\sigma \left(\frac{2}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Delta\Psi_{p,j}^{\lambda_1,\lambda_2} + 2\sqrt{\Sigma_{p,p}} \Delta\Psi_{p,p}^{\lambda_1,\lambda_2}\right) \\
&+ \sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1,\lambda_2} \\
0 =& \mathring{\Delta}_\sigma^2 \Delta\Psi_{p,p}^{\lambda_1,\lambda_2} + \mathring{\Delta}_\sigma \left(\frac{2}{\sqrt{\Sigma_{p,p}}} \sum_{j \neq p} \Sigma_{p,j} \Delta\Psi_{p,j}^{\lambda_1,\lambda_2} + \frac{2}{\sqrt{\Sigma_{p,p}}} \Sigma_{p,p} \Delta\Psi_{p,p}^{\lambda_1,\lambda_2}\right) \\
&+ \sum_{i,j} \Sigma_{i,j} \Delta\Psi_{i,j}^{\lambda_1,\lambda_2}
\end{aligned}$$

$$
\begin{aligned}
0 =& \mathring{\Delta}_\sigma^2 \Delta\Psi^{\lambda_1,\lambda_2}_{p,p} + \mathring{\Delta}_\sigma 2 \sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}} \Delta\Psi^{\lambda_1,\lambda_2}_{p,j} + \sum_{i,j} \Sigma_{i,j} \Delta\Psi^{\lambda_1,\lambda_2}_{i,j} \\
\mathring{\Delta}_\sigma =& -\frac{2 \sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}} \Delta\Psi^{\lambda_1,\lambda_2}_{p,j}}{2\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}} \\
& \pm \frac{\sqrt{\left(2 \sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}} \Delta\Psi^{\lambda_1,\lambda_2}_{p,j}\right)^2 - 4\Delta\Psi^{\lambda_1,\lambda_2}_{p,p} \sum_{i,j} \Sigma_{i,j} \Delta\Psi^{\lambda_1,\lambda_2}_{i,j}}}{2\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}} \\
\mathring{\Delta}_\sigma =& -\sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{p,j}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}} \pm \sqrt{\left(\sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{p,j}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}}\right)^2 - \sum_{i,j} \Sigma_{i,j} \frac{\Delta\Psi^{\lambda_1,\lambda_2}_{i,j}}{\Delta\Psi^{\lambda_1,\lambda_2}_{p,p}}} \qquad \text{(A.40)}
\end{aligned}
$$

□

Proof of Theorem 4.13. As a basis for a definition for the thresholds in this case, Equation 4.8 is adapted to the bivariate case with $\lambda_1 = 1$ and $\lambda_2 = 0$ in a first step in Equation A.41.

$$
\begin{aligned}
\Delta\Psi^{1,0} =& \Omega^1 + \mathrm{E}\left[w^1\right] \mathrm{E}\left[w^1\right]^\top - \Omega^0 - \mathrm{E}\left[w^0\right] \mathrm{E}\left[w^0\right]^\top \\
=& \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} - \frac{\left(1-\rho^2\right) \sigma_1^2 \sigma_2^2}{(n-3) \left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \\
& + \frac{1}{\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2} \\
& \begin{bmatrix} \left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)^2 & \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right) \\ \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right) & \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)^2 \end{bmatrix} \\
=& \frac{1}{\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2} \begin{bmatrix} \frac{1}{4}\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2 & \frac{1}{4}\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2 \\ \frac{1}{4}\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2 & \frac{1}{4}\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2 \end{bmatrix} \\
& - \frac{\left(1-\rho^2\right) \sigma_1^2 \sigma_2^2}{(n-3) \left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \\
& + \frac{1}{\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right)^2} \\
& \begin{bmatrix} \left(\sigma_2^2 - \rho\sigma_1\sigma_2\right)^2 & \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right) \\ \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)\left(\sigma_2^2 - \rho\sigma_1\sigma_2\right) & \left(\sigma_1^2 - \rho\sigma_1\sigma_2\right)^2 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{\sigma_1^2-\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}\\
&\qquad\begin{bmatrix}\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2 & \frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)\\ \frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right) & \frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)-\sigma_1^2+\rho\sigma_1\sigma_2\end{bmatrix}\\
&-\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{(n-3)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}
\end{aligned}
\tag{A.41}
$$

This result can be used to derive the term $\sum_{i,j}\Sigma_{i,j}\Delta\Psi_{i,j}^{\lambda_1,\lambda_2}$, as presented in Equation A.42.

$$
\begin{aligned}
\sum_{i,j}\Sigma_{i,j}\Delta\Psi_{i,j}^{1,0}=&\frac{1}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}\Bigg(\\
&\sigma_1^2\left(\left(\sigma_1^2-\sigma_2^2\right)\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2\right)-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\\
&+2\rho\sigma_1\sigma_2\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2+\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\\
&+\sigma_2^2\left(\left(\sigma_1^2-\sigma_2^2\right)\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)-\sigma_1^2+\rho\sigma_1\sigma_2\right)-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\Bigg)\\
=&\frac{1}{\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)^2}\Bigg(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)\\
&-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)\Bigg)\\
=&\frac{\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}
\end{aligned}
\tag{A.42}
$$

Plugging $\Psi^{1,0}$ into Theorem 4.11 gives the special case of the critical value as shown in Equation A.43.

$$
\begin{aligned}
\mathring{\Delta}_\rho=&-\frac{\sum_{i,j}\Sigma_{i,j}\Delta\Psi_{i,j}^{\lambda_1,\lambda_2}}{2\sqrt{\Sigma_{p,p}\Sigma_{q,q}}\Delta\Psi_{p,q}^{\lambda_1,\lambda_2}}\\
=&-\frac{\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{2\sigma_1\sigma_2\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2+\frac{1+\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)}
\end{aligned}
\tag{A.43}
$$

Under the assumption of a training sample of infinite size, the critical value is derived in Equation A.44.

$$\begin{aligned}\mathring{\Delta}_{\rho}^{\infty} =& \lim_{n\to\infty} -\frac{\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{2\sigma_1\sigma_2\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2+\frac{1+\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)}\\ =& -\frac{\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{2\sigma_1\sigma_2\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2}\\ =& -\frac{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}{2\sigma_1\sigma_2}\\ =&\rho-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1\sigma_2}\end{aligned} \tag{A.44}$$

□

Proof of Theorem 4.14. For $\mathring{\Delta}_{\sigma}$, additionally deriving $\sum_j \Sigma_{p,j}\Delta\Psi_{p,j}^{\lambda_1,\lambda_2}$ is required. Without loss of generality, $p=1$ can be assumed since the threshold for $p=2$ can always be calculated by using a slightly modified version of Σ.

With $d=\sigma_1^2-\sigma_2^2$ and $m=\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2$, the required $\sum_j \Sigma_{p,j}\Delta\Psi_{p,j}^{\lambda_1,\lambda_2}$ is derived in Equation A.45.

$$\begin{aligned}\sum_j \Sigma_{1,j}\Delta\Psi_{1,j}^{1,0} =&\sigma_1^2\left(\left(\sigma_1^2-\sigma_2^2\right)\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)+\sigma_2^2-\rho\sigma_1\sigma_2\right)-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\\ &+\rho\sigma_1\sigma_2\left(\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2+\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\right)\\ =&\frac{1}{4}\left(\sigma_1^2-\sigma_2^2\right)^2\left(\sigma_1^2+\rho\sigma_1\sigma_2\right)+\sigma_1^2\left(\sigma_1^2-\sigma_2^2\right)\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)\\ &-\frac{1-\rho^2}{n-3}\sigma_1^2\sigma_2^2\left(\sigma_1^2-\rho\sigma_1\sigma_2\right)\\ =&\frac{1}{4}d^2\left(\sigma_1^2+\rho\sigma_1\sigma_2\right)+\sigma_1^2 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1^2-\rho\sigma_1\sigma_2\right)\end{aligned} \tag{A.45}$$

Plugging $\sum_{i,j}\Sigma_{i,j}\Delta\Psi_{i,j}^{1,0}$ derived in Theorem 4.13 (Equation A.45) and $\sum_j \Sigma_{1,j}\Delta\Psi_{1,j}^{1,0}$ into Theorem 4.12 yields the claimed critical value as presented in Equation A.46.

$$\mathring{\Delta}_{\sigma}=-\sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}}\frac{\Delta\Psi_{p,j}^{\lambda_1,\lambda_2}}{\Delta\Psi_{p,p}^{\lambda_1,\lambda_2}}\pm\sqrt{\left(\sum_j \frac{\Sigma_{p,j}}{\sqrt{\Sigma_{p,p}}}\frac{\Delta\Psi_{p,j}^{\lambda_1,\lambda_2}}{\Delta\Psi_{p,p}^{\lambda_1,\lambda_2}}\right)^2-\sum_{i,j}\Sigma_{i,j}\frac{\Delta\Psi_{i,j}^{\lambda_1,\lambda_2}}{\Delta\Psi_{p,p}^{\lambda_1,\lambda_2}}}$$

$$
\begin{aligned}
&= -\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\\
&\quad\pm\left(\left(\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^2\right.\\
&\qquad\left.-\frac{\left(\frac{1}{4}d^2-m\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^{\frac{1}{2}}
\end{aligned}
\tag{A.46}
$$

Under the assumption of a training sample of infinite size, the critical value is derived in Equation A.47.

$$
\begin{aligned}
\mathring{\Delta}_{\sigma}^{\infty} &= \lim_{n\to\infty} -\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\\
&\quad\pm\left(\left(\frac{\frac{1}{4}d^2\left(\sigma_1+\rho\sigma_2\right)+\sigma_1 d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^2\right.\\
&\qquad\left.-\frac{\left(\frac{1}{4}d^2-m\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\frac{1}{4}d^2+d\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)-m}\right)^{\frac{1}{2}}\\
&= -\frac{\frac{1}{4}d\left(\sigma_1+\rho\sigma_2\right)+\sigma_1\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)\left(\sigma_1-\rho\sigma_2\right)}{\frac{1}{4}d+\sigma_2^2-\rho\sigma_1\sigma_2}\\
&\quad\pm\sqrt{\left(\frac{\frac{1}{4}d\left(\sigma_1+\rho\sigma_2\right)+\sigma_1\left(\sigma_2^2-\rho\sigma_1\sigma_2\right)}{\frac{1}{4}d+\sigma_2^2-\rho\sigma_1\sigma_2}\right)^2-\frac{\frac{1}{4}d\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\frac{1}{4}d+\sigma_2^2-\rho\sigma_1\sigma_2}}\\
&= -\frac{\sigma_1^3-3\rho\sigma_1^2\sigma_2+3\sigma_1\sigma_2^2-\rho\sigma_2^3}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\\
&\quad\pm\sqrt{\left(\frac{\sigma_1^3-3\rho\sigma_1^2\sigma_2+3\sigma_1\sigma_2^2-\rho\sigma_2^3}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\right)^2-\frac{\left(\sigma_1^2-\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}}\\
&= -\sigma_1-\frac{\rho\sigma_2\left(\sigma_1^2-\sigma_2^2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\\
&\quad\pm\sqrt{\left(\sigma_1+\frac{\rho\sigma_2\left(\sigma_1^2-\sigma_2^2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}\right)^2-\frac{\left(\sigma_1^2-\sigma_2^2\right)\left(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)}{\sigma_1^2+3\sigma_2^2-4\rho\sigma_1\sigma_2}}
\end{aligned}
\tag{A.47}
$$

□

Proof of Theorem 4.15. The expected out-of-sample error variance introduced in

Theorem 4.5 can be differentiated with respect to λ, resulting in Equation A.48.

$$\begin{aligned}\frac{\delta Var}{\delta \lambda} =& \sum_{i,j} \tilde{\Sigma}_{i,j} \left(2\left(\lambda - 1\right)\Omega_{i,j}\right) \\ &+ \sum_{i,j} \tilde{\Sigma}_{i,j} \left(\frac{2\lambda}{k^2} + \left(\mathrm{E}\left[w_i^O\right] - \mathrm{E}\left[w_j^O\right]\right)\frac{1-2\lambda}{k} + 2\left(\lambda - 1\right)\mathrm{E}\left[w_i^O\right]\mathrm{E}\left[w_j^O\right]\right)\end{aligned} \tag{A.48}$$

Solving $\frac{\delta Var}{\delta \mathring{\lambda}} = 0$ for $\mathring{\lambda}$ then quickly results in the critical value in Equation A.49.

$$\mathring{\lambda} = \frac{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(\Omega_{i,j} - \frac{1}{2k}\mathrm{E}\left[\hat{w}_i^O\right] - \frac{1}{2k}\mathrm{E}\left[\hat{w}_j^O\right] + \mathrm{E}\left[\hat{w}_i^O\right]\mathrm{E}\left[\hat{w}_j^O\right]\right)}{\sum_{i,j} \tilde{\Sigma}_{i,j} \left(\Omega_{i,j} + \left(\frac{1}{k}\right)^2 - \frac{1}{k}\mathrm{E}\left[\hat{w}_i^O\right] - \frac{1}{k}\mathrm{E}\left[\hat{w}_j^O\right] + \mathrm{E}\left[\hat{w}_i^O\right]\mathrm{E}\left[\hat{w}_j^O\right]\right)} \tag{A.49}$$

The second derivative of the expected out-of-sample error variance with respect to λ is shown in Equation A.50.

$$\begin{aligned}\frac{\delta Var}{\delta^2 \lambda} =& \sum_{i,j} \tilde{\Sigma}_{i,j} \left(2\Omega_{i,j} + \frac{2}{k^2} - \left(\mathrm{E}\left[w_i^O\right] - \mathrm{E}\left[w_j^O\right]\right)\frac{2}{k} + 2\mathrm{E}\left[w_i^O\right]\mathrm{E}\left[w_j^O\right]\right) \\ =& \sum_{i,j} \tilde{\Sigma}_{i,j} \left(2\Omega_{i,j} + 2\left(\frac{1}{k} - \mathrm{E}\left[w_i^O\right]\right)\left(\frac{1}{k} - \mathrm{E}\left[w_j^O\right]\right)\right) \\ =& 2\vec{1}_k^\top \left(\Omega\tilde{\Sigma}\right)\vec{1}_k + 2\left(\frac{1}{k} - \mathrm{E}\left[w^O\right]\right)^\top \tilde{\Sigma}\left(\frac{1}{k} - \mathrm{E}\left[w^O\right]\right)\end{aligned} \tag{A.50}$$

Since both Ω and $\tilde{\Sigma}$ are covariance matrices, $\Omega\tilde{\Sigma}$ is a product of positive definite matrices, which is again a positive definite matrix. Consequently, the first term of the combined error variance in Theorem 4.5 is positive. Likewise, since $\tilde{\Sigma}$ is positive definite, the second term is positive. Overall, the second derivative is always positive and $\mathring{\lambda}$ minimizes the combined error variance. □

References

Aiolfi, M. and A. Timmermann (2005). Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics 135*, 31 – 53.

Aksu, C. and S. Gunter (1992). An Empirical Analysis of the Accuracy of SA, OLS, ERLS and NRLS Combination Forecasts. *International Journal of Forecasting 8*(1), 27 – 43.

Andreassen, P. B. (1988). Explaining the Price-Volume Relationship: The Difference Between Price Changes and Changing Prices. *Organizational Behavior and Human Decision Processes 41*, 371 – 389.

Armstrong, J. (2001). Combining Forecasts. In *Principles of Forecasting*, pp. 417 – 439. Springer.

Armstrong, J. S. and F. Collopy (1998). Integration of Statistical Methods and Judgment for Time Series Forecasting: Principles From Empirical Research. In G. Wright and P. Goodwin (Eds.), *Forecasting with Judgment*, pp. 269 – 293. John Wiley & Sons Ltd.

Ashley, R. (1985). On the Optimal Use of Suboptimal Forecasts of Explanatory Variables. *Journal of Business & Economic Statistics 3*(2), 129 – 131.

Basu, S. and S. Markov (2004). Loss Function Assumptions in Rational Expectations Tests on Financial Analysts' Earnings Forecasts. *Journal of Accounting and Economics 38*, 171 – 203.

Bates, J. and C. Granger (1969). The Combination of Forecasts. *Journal of the Operational Research Society 20*(4), 451 – 468.

Bhandari, G., K. H., and R. Deaves (2008). Debiasing Investors with Decision Support Systems: An Experimental Investigation. *Decision Support Systems 46*(1), 399 – 410.

Blanc, S. and P. Ruchser (2016). Robust Debiasing of Judgmental Forecasts with Structural Changes. In *Multikonferenz Wirtschaftsinformatik (MKWI) 2016*, Volume 2, pp. 1193 – 1204. Universitätsverlag Ilmenau.

Blanc, S. and T. Setzer (2015a). Improving Forecast Accuracy by Guided Manual Overwrite in Forecast Debiasing. In *Twenty-Third European Conference on Information Systems (ECIS)*.

Blanc, S. and T. Setzer (2016a). Bias–Variance-Aware and Robust Shrinkage of Weights in Forecast Combination. *Management Science*, under review.

Blanc, S. and T. Setzer (2016b). When to Choose the Simple Average in Forecast Combination. *Journal of Business Research 69*(10), 3951 – 3962.

Blanc, S. M. and T. Setzer (2015b). Analytical Debiasing of Corporate Cash Flow Forecasts. *European Journal of Operational Research 243*(3), 1004 – 1015.

Blattberg, R. C. and S. J. Hoch (1990). Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science 36*(8), 887 – 899.

Bohara, A., R. McNown, and J. Batts (1987). A Re-Evaluation of the Combination and Adjustment of Forecasts. *Applied Economics 19*(4), 437 – 445.

Bohrnstedt, G. W. and A. S. Goldberger (1969). On the Exact Covariance of Products of Random Variables. *Journal of the American Statistical Association 64*(328), 1439 – 1442.

Boudt, K., P. Goeij, J. Thewissen, and G. Van Campenhout (2014). Analysts' Forecast Error: A Robust Prediction Model and Its Short-Term Trading Profitability. *Accounting & Finance 55*(3), 683 – 715.

Bowley, A. (1901). *Elements of Statistics*. Scribner's, New York.

Bowman, E. (1963). Consistency and Optimality in Managerial Decision Making. *Management Science 9*(2), 310 – 321.

Brandon, C., R. Fritz, and J. Xander (1983). Econometric Forecasts: Evaluation and Revision. *Applied Economics 15*(2), 187 – 201.

Brockwell, P. and R. Davis (2013). *Time Series: Theory and Methods*. Springer.

Budescu, D. and E. Chen (2015). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science 61*(2), 267 – 280.

Bunn, D. and G. Wright (1991). Interaction of Judgemental and Statistical Forecasting Methods: Issues & Analysis. *Management Science 37*(5), 501 – 518.

Bunn, D. W. (1985). Statistical Efficiency in the Linear Combination of Forecasts. *International Journal of Forecasting 1*(2), 151 – 163.

Chatterjee, S. and A. Hadi (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science 1*(3), 379 – 393.

Claeskens, G., J. Magnus, A. Vasnev, and W. Wang (2016). The Forecast Combination Puzzle: A Simple Theoretical Explanation. *International Journal of Forecasting 32*(3), 754 – 762.

Clemen, R. (1989). Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting 5*(4), 559 – 583.

Clemen, R. and R. Winkler (1986). Combining Economic Forecasts. *Journal of Business & Economic Statistics 4*(1), 39 – 46.

Cleveland, R., W. Cleveland, J. McRae, and I. Terpenning (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics 6*(1), 3 – 73.

Cook, R. (1977). Detection of Influential Observation in Linear Regression. *Technometrics 19*(1), 15 – 18.

Cryer, J. and K. Chan (2008). *Time Series Analysis: With Applications in R*. Springer.

Davis-Stober, C., D. Budescu, J. Dana, and S. Broomell (2014). When is a Crowd Wise? *Decision 1*(2), 79 – 101.

De Gooijer, J. and R. Hyndman (2006). 25 years of time series forecasting. *International Journal of Forecasting 22*(3), 443 – 473.

de Menezes, L. M., D. W. Bunn, and J. W. Taylor (2000). Review of Guidelines for the Use of Combined Forecasts. *European Journal of Operational Research 120*(1), 190 – 204.

Dickinson, J. P. (1973). Some Statistical Results in the Combination of Forecasts. *Journal of the Operational Research Society 24*(2), 253 – 260.

Diebold, F. and J. Lopez (1996). Forecast Evaluation and Combination. In M. G.S. and R. C.R. (Eds.), *Statistical Methods in Finance*, Volume 14 of *Handbook of Statistics*, pp. 241 – 268. Elsevier.

Diebold, F. and P. Pauly (1987). Structural Change and the Combination of Forecasts. *Journal of Forecasting 6*(1), 21 – 40.

Diebold, F. and P. Pauly (1990). The Use of Prior Information in Forecast Combination. *International Journal of Forecasting 6*(4), 503 – 508.

Edmundson, B., M. Lawrence, and M. O'Connor (1988). The Use of Non-Time Series Information in Sales Forecasting: A Case Study. *Journal of Forecasting 7*(3), 201 – 211.

Eggleton, I. R. C. (1982). Intuitive time series extrapolation. *Journal of Accounting Research 20*, 68 – 102.

Elgers, P. T., M. H. Lo, and D. Murray (1995). Note on Adjustments to Analysts' Earnings Forecasts Based Upon Systematic Crosssectional Components of Prior-period Errors. *Management Science 41*(8), 1392 – 1396.

Elliott, G. (2011). Averaging and the Optimal Combination of Forecasts. Technical report, University of California, San Diego.

Enns, S. T. (2002). MRP Performance Effects due to Forecast Bias and Demand Uncertainty. *European Journal of Operational Research 138*(1), 87 – 102.

Figlewski, S. and T. Urich (1983). Optimal Aggregation of Money Supply Forecasts: Accuracy, Profitability and Market Efficiency. *The Journal of Finance 38*(3), 695 – 710.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos (2009). Effective Forecasting and Judgmental Adjustments: An Empirical Evaluation and Strategies for Improvement in Supply-Chain Planning. *International Journal of Forecasting 25*(1), 3 – 23.

Franses, P. (2008). Merging Models and Experts. *International Journal of Forecasting 24*(1), 31 – 33.

Franses, P. (2013). Improving Judgmental Adjustment of Model-Based Forecasts. *Mathematics and Computers in Simulation 93*, 1 – 8.

Franses, P. and R. Legerstee (2010). Do Experts' Adjustments on Model-Based SKU-Level Forecasts Improve Forecast Quality? *Journal of Forecasting 29*(3), 331 – 340.

Franses, P. H. (2011). Averaging Model Forecasts and Expert Forecasts: Why Does It Work? *Interfaces 41*(2), 177 – 181.

Gardner, E. (2006). Exponential Smoothing: The State of the Art - Part II. *International Journal of Forecasting 22*(4), 637 – 666.

Gardner, E. and E. McKenzie (2011). Why the Damped Trend Works. *Journal of the Operational Research Society 62*(6), 1177 – 1180.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining Expert Forecasts: Can Anything Beat the Simple Average? *International Journal of Forecasting 29*(1), 108 – 121.

George, J. F., K. Duffy, and M. Ahuja (2000). Countering the Anchoring and Adjustment Bias with Decision Support Systems. *Decisions Support Systems 29*(2), 195 – 206.

Gigone, D. and R. Hastie (1997). Proper Analysis of the Accuracy of Group Judgments. *Psychological Bulletin 121*(1), 149 – 167.

Gilchrist, W. (1974). Statistical forecasting - The state of the art. *Omega - The International Journal of Management Science 2*(6), 733 – 750.

Giloni, A., J. Simonoff, and B. Sengupta (2006). Robust Weighted LAD Regression. *Computational Statistics & Data Analysis 50*(11), 3124 – 3140.

Goodwin, P. (1996). Statistical Correction of Judgmental Point Forecasts and Decisions. *Omega 24*(5), 551 – 559.

Goodwin, P. (1997). Adjusting Judgemental Extrapolations using Theil's Method and Discounted Weighted Regression. *Journal of Forecasting 16*, 37 – 46.

Goodwin, P. (2000). Correct or Combine? Mechanically Integrating Judgmental Forecasts With Statistical Methods. *International Journal of Forecasting 16*, 261 – 275.

Goodwin, P. (2002). Integrating Management Judgment and Statistical Methods to Improve Short-Term Forecasts. *Omega 30*(2), 127 – 135.

Goodwin, P. and G. Wright (1994). Heuristics, Biases and Improvement Strategies in Judgmental Time Series Forecasting. *Omega 22*(6), 553 – 568.

Gormley, F. M. and N. Meade (2007). The Utility of Cash Flow Forecasts in the Management of Corporate Cash Balances. *European Journal of Operational Research 182*(2), 923 – 935.

Graham, J. R. and C. R. Harvey (2001). The Theory and Practice of Corporate Finance: Evidence from the Field. *Journal of Financial Economics 60*(2), 187 – 243.

Granger, C. and P. Newbold (2014). *Forecasting Economic Time Series*. Academic Press.

Granger, C. W. and R. Ramanathan (1984). Improved Methods of Combining Forecasts. *Journal of Forecasting 3*(2), 197 – 204.

Groeneveld, R. (1991). An Influence Function Approach to Describing the Skewness of a Distribution. *The American Statistician 45*(2), 97 – 102.

Gupta, A. K. and D. K. Nagar (1999). *Matrix Variate Distributions*. CRC Press.

Gupta, S. and P. Wilton (1987). Combination of Forecasts: An Extension. *Management Science 33*(3), 356 – 372.

Harrison, J. and M. West (1999). *Bayesian Forecasting and Dynamic Models*. Springer.

Harvey, N. (1995). Why are Judgements Less Consistent in Less Predictable Task Situations? *Organizational Behavior and Human Decision Processes 63*, 247 – 263.

Hastie, R. (1986). Experimental Evidence on Group Accuracy. *Decision Research 2*, 129 – 157.

Hastie, R. and T. Kameda (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review 112*(2), 494 – 508.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2 ed.). Springer.

Hendry, D. and M. Clements (2004). Pooling of Forecasts. *The Econometrics Journal 7*(1), 1 – 31.

Hill, G. (1982). Group Versus Individual Performance: Are N+1 Heads Better Than One? *Psychological Bulletin 91*(3), 517 – 539.

Hogarth, R. M. and S. Makridakis (1981). Forecasting and Planning: An Evaluation. *Management Science 27*(2), 115 – 138.

Hyndman, R. (2015). *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 6.1.

Hyndman, R. and Y. Khandakar (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software 27*(1), 1 – 22.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Kang, H. (1986). Unstable Weights in the Combination of Forecasts. *Management Science 32*(6), 683 – 695.

Kerr, N. and R. Tindale (2011). Group-Based Forecasting?: A Social Psychological Analysis. *International Journal of Forecasting 27*(1), 14 – 40.

Kim, C., D. Mauer, and A. Sherman (1998). The Determinants of Corporate Liquidity: Theory and Evidence. *Journal of Financial and Quantitative Analysis 33*(3), 335 – 359.

Klassen, R. and B. Flores (2001). Forecasting Practices of Canadian Firms: Survey Results and Comparisons. *International Journal of Production Economics 70*(2), 163 – 174.

Kleinmuntz, B. (1990). Why We Still Use Our Heads Instead of Formulas: Toward an Integrative Approach. *Psychological Bulletin 107*(3), 296 – 310.

Koenker, R. (2015). *quantreg: Quantile Regression*. R package version 5.19.

Kshirsagar, A. M. (1961). Some Extensions of the Multivariate t-Distribution and the Multivariate Generalization of the Distribution of the Regression Coefficient. *57*(1), 80 – 85.

Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *Journal of Econometrics 54*(1), 159 – 178.

Lawrence, M., R. Edmundson, and M. O'Connor (1985). An Examination of the Accuracy of Judgmental Extrapolation of Time Series. *International Journal of Forecasting 1*(1), 25 – 35.

Lawrence, M., P. Goodwin, M. O'Connor, and D. Oenkal (2006). Judgmental Forecasting: A Review of Progress Over the Last 25 Years. *International Journal of Forecasting 22*, 493 – 618.

Lawrence, M. and M. O'Connor (1995). The Anchor and Adjustment Heuristic in Time-Series Forecasting. *Journal of Forecasting 14*(5), 443 – 451.

Lawrence, M. and M. O'Connor (1996). Judgement or Models: The Importance of Task Differences. *Omega 24*(3), 245 – 254.

Lawrence, M., M. O'Connor, and B. Edmundson (2000). A Field Study of Sales Forecasting Accuracy and Processes. *European Journal of Operational Research 122*(1), 151 – 160.

Lawrence, M. J., R. H. Edmundson, and M. J. O'Connor (1986). The Accuracy of Combining Judgemental and Statistical Forecasts. *Management Science 32*(12), 1521 – 1532.

Leitner, J. and U. Leopold-Wildburger (2011). Experiments on Forecasting Behavior with Several Sources of Information - A Review of the Literature. *European Journal of Operational Research 213*(3), 459 – 469.

Lim, J. S. and M. O'Connor (1996). Judgmental Forecasting with Interactive Forecasting Support Systems. *Decision Support Systems 16*(4), 339 – 357.

llmakunnas, P. (1990). Forecast Pretesting and Correction. *Journal of Business & Economic Statistics 8*(4), 475 – 480.

Lobo, G. and R. Nair (1990). Combining Judgmental and Statistical Forecasts: An Application to Earnings Forecasts. *Decision Sciences 21*(2), 446 – 460.

Lopes, L. L. and G. C. Oden (1987). Distinguishing Between Random and Non-Random Events. *Journal of Experimental Psychology 13*, 392 – 400.

Makridakis, S. (1988). Metaforecasting: Ways of Improving Forecasting Accuracy and Usefulness. *International Journal of Forecasting 4*(3), 467 – 491.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition. *Journal of Forecasting 1*(2), 111 – 153.

Makridakis, S. and M. Hibon (2000). The M3-Competition: Results, Conclusions and Implications. *International Journal of Forecasting 16*(4), 451 – 476.

Makridakis, S., S. Wheelwright, and R. Hyndman (2008). *Forecasting Methods and Applications*. John Wiley & Sons.

Mannes, A., J. Soll, and R. Larrick (2014). The Wisdom of Select Crowds. *Journal of Personality and Social Psychology, 107*(2), 276 – 299.

McCarthy, T., D. Davis, S. Golicic, and J. Mentzer (2006). The Evolution of Sales Forecasting Management: A 20-year Longitudinal Study of Forecasting Practices. *Journal of Forecasting 25*(5), 303 – 324.

Miller, C., R. Clemen, and R. Winkler (1992). The Effect of Nonstationarity on Combined Forecasts. *International Journal of Forecasting 7*(4), 515 – 529.

Moriarty, M. M. (1985). Design Features of Forecasting Systems Involving Management Judgements. *Journal of Marketing Research 22*, 353 – 364.

O'Connor, M., W. Remus, and K. Griggs (1993). Judgmental forecasting in times of change. *International Journal of Forecasting 9*, 163 – 172.

Pearson, K. (1926). Researches on the Mode of Distribution of the Constants of Samples Taken at Random From a Bivariate Normal Population. *Proceedings of the Royal Society of London. Series A 112*(760), 1 – 14.

Pretis, F., L. Schneider, J. Smerdon, and D. Hendry (2016). Detecting Volcanic Eruptions in Temperature Reconstructions by Designed Break-Indicator Saturation. *Journal of Economic Surveys 30*(3), 403 – 429.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reid, D. (1968). Combining Three Estimates of Gross Domestic Product. *Economica 35*(140), 431 – 444.

Reimers, S. and N. Harvey (2011). Sensitivity to Autocorrelation in Judgmental Time Series Forecasting. *International Journal of Forecasting 27*(4), 1196 – 1214.

Remus, W. and J. Kottemann (1995). Anchor-and-Adjustment Behaviour in a Dynamic Decision Environment. *Decision Support Systems 15*(1), 63 – 74.

Romanovsky, V. I. (1926). On the Distribution of the Regression Coefficient in Samples From Normal Population. *(Bulletin de l'Académie des Sciences de l'URSS 20*(9), 643 – 648.

Rousseeuw, P. and A. Leroy (2005). *Robust Regression and Outlier Detection*. John Wiley & Sons.

Sanders, N. and K. Manrodt (2003). The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice. *Omega 31*(6), 511 – 522.

Sanders, N. and L. Ritzman (1990). Improving Short-Term Forecasts. *Omega 18*(4), 365 – 373.

Sanders, N. and L. Ritzman (1995). Bringing Judgment Into Combination Forecasts. *Journal of Operations Management 13*(4), 311 – 321.

Sanders, N. and L. Ritzman (2001). Judgmental Adjustment of Statistical Forecasts. In *Principles of Forecasting*, pp. 405 – 416. Springer.

Sanders, N. R. and L. P. Ritzman (2004). Integrating Judgmental and Quantitative Forecasts: Methodologies for Pooling Marketing and Operations Information. *International Journal of Operations & Production Management 24*(5), 514 – 529.

Schmittlein, D., J. Kim, and D. Morrison (1990). Combining Forecasts: Operational Adjustments to Theoretically Optimal Rules. *Management Science 36*(9), 1044 – 1056.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461 – 464.

Shaffer, S. (1998). Information Content of Forecast Errors. *Economics Letters 59*, 45 – 48.

Smith, J. and K. Wallis (2009). A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics 71*(3), 331 – 355.

Stewart, T. and P. Reagan-Cirincione (1991). Coefficients for Debiasing Forecasts. *Monthly Weather Review 119*(8), 2047 – 2051.

Stock, J. and M. Watson (1999). A Comparison of Linear and Nonlinear Models for Forecasting Macroeconomic Time Series. In R. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting*, pp. 1 – 44. Oxford University Press.

Stock, J. and M. Watson (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting 23*(6), 405 – 430.

Syntetos, A., K. Nikolopoulos, J. Boylan, R. Fildes, and P. Goodwin (2009). The Effects of Integrating Management Judgement Into Intermittent Demand Forecasts. *International Journal of Production Economics 118*(1), 72 – 81.

Theil, H. (1966). *Applied Economic Forecasting*. North Holland Publishing Company.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267 – 288.

Timmermann, A. (2006). Forecast Combinations. *Handbook of Economic Forecasting*, 135 – 196.

Utley, J. (2011). Removing Systematic Bias From Demand Forecasts. In K. D. Lawrence and R. K. Klimberg (Eds.), *Advances in Business and Management Forecasting, Volume 8*, pp. 3 – 12. Emerald Group Publishing.

Wallsten, T., D. Budescu, I. Erev, and A. Diederich (1997). Evaluating and Combining Subjective Probability Estimates. *Journal of Behavioral Decision Making 10*(3), 243 – 268.

Webby, R. and M. O'Connor (1996). Judgemental and Statistical Time Series Forecasting: A Review of the Literature. *International Journal of Forecasting 12*(1), 91 – 118.

Winkler, R. L. and R. T. Clemen (1992). Sensitivity of Weights in Combining Forecasts. *Operations Research 40*(3), 609 – 614.

Yule, G. (1926). Why Do We Sometimes Get Nonsense-Correlations Between Time-Series? – A Study in Sampling and the Nature of Time-Series. *Journal of the Royal Statistical Society 89*(1), 1 – 63.

Zeileis, A., C. Kleiber, W. Kraemer, and K. Hornik (2003). Testing and Dating of Structural Changes in Practice. *Computational Statistics & Data Analysis 44*(1), 109 – 123.

Zeileis, A., F. Leisch, K. Hornik, and C. Kleiber (2002). strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software 7*(2), 1 – 38.