2019

Forecast Combination with Multiple Models and Expert Correlations

David P. Soule
Virginia Commonwealth University

Forecast Combination with Multiple Models and Expert Correlations

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

DAVID PATTERSON SOULE

MBA New York University - 1981

BS Mechanical Engineering Massachusetts Institute of Technology - 1979

Director: Prof. Jason Merrick,

PhD. Program in Systems Modeling and Analysis

Virginia Commonwealth University

Richmond, Virginia

May, 2019

Acknowledgements

I started this journey nine years ago, when I took a VCU class to see how I liked going back to school. During this lengthy time, my wife and family have been lovingly supportive and understanding of the time spent away from the family taking evening classes or down in my office doing school work. In addition, my wife volunteered and did a great job helping to edit and proofread my dissertation. Without their support I could never have completed this journey.

My employer, Capital One, and two executives Patrick Gemmell and Dave Lewis, were key to making this journey possible. Capital One fully funded my tuition as an employee benefit. Patrick and Dave were highly supportive and encouraged me, as well as, provided the flexibility for me to take time away from work to attend day classes. While Super-D, Donna Tees, did a great job juggling my work schedule to allow me to attend class and get my Capital One work done. Without their support, and many others at Capital One, I could never have completed this journey.

My advisor and coach, Professor Jason Merrick, was a wonderful guide for me along this journey. He initially advised me on the decision to embark on my PhD as a way to transition from business to an academic role. As my dissertation advisor, Jason, provided a unique combination of cheerfulness and encouragement with intellectual challenge and guidance. He found a good balance of providing higher level questions and direction while also partnering on some of the more confounding analytical issues we ran into. Throughout the process he actively coached and assisted with my career transition into academia. I have learned much more from Jason than how to approach a dissertation. I am very lucky to have had Jason as my advisor.

Professor Yael Gruska-Cockayne signed on as the external committee member and later signed on as a co-author of a paper coming out of this research. She has

most generously shared of her time and essentially served as a co-advisor even though she is not a VCU faculty member. She has patiently taught me how to write for an academic journal. I appreciate all her questions, feedback, and time she has shared as it has always improved my thinking and the dissertation.

In addition to Professors Merrick and Gruska-Cockayne, I have benefited from a strong committee including Dean Montserrat Fuentes, Professors QiQi Lu, and David Edwards. Previous members of the committee have also included Professors Qin Wang, Qiong Zhang. Each committee member has spent time with me and provided unique insights and ideas to make this work better.

This work has benefited from real world data shared from two sources. Focus Economics, and specifically Michael Boydell and Daniel Rempe, provided a diverse set of economic forecast data to validate our proposals with. Professors Soll, Mannes, and Larrick shared a nicely compiled set of experimental data which provided a second perspective on the validation. Having this data enhanced the strength of our findings.

From all of the above it is clear I have been blessed by the Lord Jesus Christ. He has walked with me on each step of this journey and shown me ways to move forward when I have hit walls, or shown me errors when I have moved too fast. Only with all of this support, have I been able to complete this dissertation. I am truly thankful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

FORECAST COMBINATION WITH MULTIPLE MODELS AND EXPERT CORRELATIONS

By David Patterson Soule

A submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2019.

Director: Prof. Jason Merrick,

, PhD. Program in Systems Modeling and Analysis

Combining multiple forecasts in order to generate a single, more accurate one is a well-known approach. A simple average of forecasts has been found to be robust despite theoretically better approaches, increasing availability in the number of expert forecasts, and improved computational capabilities. The dominance of a simple average is related to the small sample sizes and to the estimation errors associated with more complex methods. We study the role that expert correlation, multiple experts, and their relative forecasting accuracy have on the weight estimation error distribution. The distributions we find are used to identify the conditions when a decision maker can confidently estimate weights versus using a simple average. We also propose an improved expert weighting approach that is less sensitive to covariance estimation error while providing much of the benefit from a covariance optimal weight. These two improvements create a new heuristic for better forecast aggregation that is simple to use. This heuristic appears new to the literature and is shown to perform better than a simple average in a simulation study and by application to

economic forecast data.

CHAPTER 1

INTRODUCTION

Since Bates and Granger's (1969) seminal paper on the combination of forecasts, there has been prolific research to determine the error minimizing ways to combine forecasts. Despite many different approaches including various weighting schemes, dynamic weights, Bayesian techniques, PCA techniques, LASSO regression, and others, a simple average of the component forecasts frequently dominates more complex techniques (Clemen, 1986; Conflitti et al., 2015; Diebold, 1989; Genre et al., 2013; Makridakis, 1982; Stock and Watson, 2004). Decision makers continue to use a simple average because it is simpler than more complex methods and has proven to be equally, if not more accurate, in practice.

Researchers have asked, "Can anything beat the simple average?"(Clemen, 1989; Genre et al., 2013). This phenomena has been named the "Forecast Combination Puzzle" and has initiated a more focused series of investigations (Blanc and Setzer, 2016; Claeskens et al., 2016; Hsiao and Wan, 2014; Schmittlein et al., 1990; Smith and Wallis, 2009; Stock and Watson, 1998). This work has increased our understanding of the Puzzle, but has primarily focused on two-forecast systems to keep the analysis tractable. Recently, Mannes et al. (2014) have shown that rather than estimating weights, simply taking an average of the top five experts frequently can out-perform a simple average of all experts. This is an exciting empirical result, but the theory of why this is true is not clear.

In this dissertation, we take a different approach to resolving the Forecast Combination Puzzle by developing models and distributions for estimated weights in a

multi-expert combination with expert error covariances. Our approach is different from past efforts in two ways:

- We develop a model for, and simulate, the distribution of weight estimation errors with expert correlations in a multi-expert system as opposed to a two-expert combination. We show that the behavior of a less constrained, multi-expert combination is different than that of a two-expert combination.

- We use this model to develop two new weight estimation heuristics that reduce the estimation error in existing, simpler forecast combination methods rather than attempting to develop a new, more complex method

The heuristics developed in this dissertation are tested in simulation studies and in empirical studies using 74 unique sets of historical economic forecasts from multiple countries and on 23 expert estimation data sets. The empirical investigation uses one of the more diverse sets of forecasts to be found in a forecast combination study.

This work results in three new contributions to the forecast combination body of knowledge:

1. We propose an improved method to estimate forecast combination weights that is demonstrated to beat a simple average (and in many cases, an average of the top 5 experts) on a statistically significant basis across a wide range of forecast data. To the author's knowledge, this is the first time that using estimated weights to combine forecasts have been found, in practice across a large empirical data set, to be significantly better, statistically, than a simple average.

2. We propose a new method to decide when an estimated weight to combine forecasts is likely to be more accurate than a simple average. This method

is demonstrated to be effective in simulations studies and on real data with sufficient estimation sample.

3. We propose an explanation for why the simple average of the top 5 experts is able to beat a simple average.

This work provides decision makers with a relatively simple approach to combine forecasts that is more accurate than a simple average. During this dissertation we have had a data sharing relationship with an economic forecasting firm that provided the historical economic forecast data. Their interest and support through data sharing demonstrates the business value of this work.

This work also provides researchers with a new model for analyzing forecast combinations that can be extended to more complex scenarios and used to develop further improvements to forecast combinations. Additional avenues of research are proposed.

1.1 Combination of multiple forecasts

The Survey of Professional Forecasters and other similar expert surveys typically use an average of as many as 30+ forecasters to form one consensus forecast for decision makers (Croushore, 1993). This wisdom of crowds approach is always as good as the average forecaster and is often better than even the most accurate individual forecaster (Bates and Granger, 1969; Mannes et al., 2014). The consensus approach across a pool of experts raises questions: Is it really optimal to treat every forecaster equally? Also, is it really optimal to include every forecaster in the consensus pool?

Work on the Forecast Combination Puzzle has found that there is a trade-off between the prediction error of using equal weights (a simple average) as opposed to estimation error of weights which optimize historical forecast combination per-

formance (Blanc and Setzer, 2016; Claeskens et al., 2016; Hsiao and Wan, 2014; Schmittlein et al., 1990; Smith and Wallis, 2009). If decision makers are unsure who is the more accurate expert and to what degree the expert is better (or worse), then they are better off not even trying to differentially weight one expert over another in the combination (Armstrong, 2001). Equal weights, in this case, will be a safer and more robust choice, as mis-assessment of each expert's performance will adversely impact the overall combination accuracy. Alternatively, if decision makers have a high degree of confidence that they know which expert is more accurate, they can consider ways to more heavily weight that expert in the forecast combination. Past work on the Forecast Combination Puzzle has focused on balancing the expected error from estimating weights based on historical performance with the expected prediction error from using equal weights across all experts in the combination. The past approaches have only considered estimating weights for all experts or using equal weights across all experts. Yet, the historical forecast accuracy of one or more experts may stand out (favorably or unfavorably) from the crowd while the accuracy of others may be relatively indistinguishable from each other. A decision maker would not have to differentiate every expert if there were a good test to determine who were differentially more (or less) accurate.

Similarly, work to find the best subset of forecasters requires the decision maker to assess the performance of each expert. Researchers have explored taking a simple average of only the most accurate few forecasters, where five has often been recommended as the best subset size (Armstrong, 2001; Clemen and Winkler, 1985, 1999; Makridakis, 1982; Mannes et al., 2014). Trimming of forecast outliers has also been shown as a method to improve upon a simple average through differential exclusion (Jose and Winkler, 2008). When including only the top five experts, is the sixth expert really that much worse than the fifth? Combining the top five forecasters,

and similar methods, all assume to some degree that the decision maker knows the relative performance of the experts. How would these approaches change if instead the decision maker knew the level of certainty that experts were different or not?

Both differentially weighting and differentially including an expert in a forecast combination requires an understanding of how confident a decision maker can be in this decision. In Chapter 3, we define expert skill and relative skill ratio to provide a framework for assessing the confidence in a weighting decision. This decision is based on the distribution of an estimated weight and the decision maker's assessment of the expert's skill. There are differing proposals for the distribution of estimated weights in a two-expert combination (Dickinson, 1973; Winkler and Clemen, 1992). This conflict is resolved by simulating the weight estimation for various skill levels, and sample sizes. The weight distributions are then used to define a critical skill ratio above which, an expert can be confident in the weighting decision. In Chapter 4, we extend the analysis of weight distributions and critical skill ratios to multi-expert combinations and develop new insights on the impact of increasing number of experts. The use of critical skill ratios based on a desired level of confidence is proposed to facilitate the decision of when to estimate weights. This approach is new to forecast combinations and can be implemented with a simple table of critical values so it meets a decision maker's desire for simplicity.

1.2 Combination of multiple forecasts with expert correlations

Frequently experts will use the same sources of information as well as similar models, these commonalities can result in expert forecasts that are highly correlated (Winkler, 1981). In addition to information and model sharing, the common absence of information, such as not foreseeing a major event or other sources of common noise, can create the appearance of correlated experts. In the case of the economic forecasts

used in this dissertation, the median correlation of an expert pair in the forecast data set ranged from $0.32 - 0.98$. In combinations of multiple experts, the likelihood and impacts of correlation between experts will increase as the number of experts and expert pairs increase.

The impact of expert correlations due to shared information and models has been well researched with a focus on the more tractable two-expert combination (Bates and Granger, 1969; Bunn, 1985; Gunter, 1992; Winkler and Clemen, 1992). The optimal weight values in this case are determined by the covariance matrix of forecast errors between experts. This covariance matrix is typically estimated based on a historical sample. In the case of multiple experts, the number of covariance parameters increases significantly and the parameter estimation error previously cited as part of the Forecast Combination Puzzle reduces the effectiveness of this approach. A second impact of expert correlations is that the wisdom of the crowd will be reduced by the degree that experts in the crowd all say very similar things (Clemen and Winkler, 1985). In this case, the size of the crowd may shout out the few more accurate experts. Forecast combinations would benefit from a method to account for expert correlation without incurring the degree of error from estimating individual expert pair correlations.

In Chapter 5 we extend our multi-expert model to analyze the impact that expert correlation has on the weight estimation process. Past studies on the sensitivity of weights have focused on two expert systems (Blanc and Setzer, 2016; Winkler and Clemen, 1992). Based on insights from the revised model, we propose only estimating one common level of expert correlation for all expert pairs, versus estimating a unique level of correlation for each expert pair.

1.3 Forecast Combination simulation study

In a simulation study (Chapter 5, Section 4) we compare different fixed and skilled-based weighting heuristics selected from the literature with our proposed critical skill ratio approaches and common correlation weight estimation method (Table 1). We find that the reduction in estimation errors when assuming a common level of correlation across all expert pairs, more than offsets the bias errors of the assumption. The highest frequency of lower errors than a simple average was achieved with an exogenous assumption of 0.3 for the common correlation level. This assumption is no more radical than just assuming fixed weights in a simple average or assuming no expert correlation in variance based weights; however it captures many of the benefits of considering expert covariance in the weight estimation process. We further find that using critical skill ratios to determine when to estimate the proposed common correlation weights (CCR) and when to use a simple average reduces the frequency that the forecast combination performance is worse than a simple average with little loss in overall Mean Average Percentage Error (MAPE) performance. This result suggests that the risk of a poor forecast can be greatly reduced by using the critical skill ratio as a decision rule.

1.4 Performance of new methods on real data

We have cited many approaches from the literature that have not withstood the test of application to real forecast data. This is the essence of the forecast combination puzzle. In Chapter 6, we demonstrate the effectiveness of the two, best performing new methods, found in the simulation study, on real data. An empirical comparison of these two methods and several benchmark methods is made using a diverse data set of historical economic forecasts for 7 different economic indicators from 15 different

	All experts	**Select experts**
Fixed weights	- Simple average (Clemen, 1989)	- Top 5 experts (Mannes et al., 2014) **- Drop experts below a critical threshold**
Skill-based weights	-Variance based weights (Bunn, 1985) - Covariance based weights (Bates and Granger, 1969) **- Common correlation based weights**	**- Drop experts below a critical threshold**
Select method, fixed or skill-based weights	- Select method based on Akaike information criterion (Schmittlein et al., 1990) **- Select method using a critical threshold**	**- Select method for each expert based on a critical threshold**

Table 1.: Forecast combination methods evaluated in simulation study. Bold faced items are new contributions from this dissertation.

countries. In addition, a second empirical comparison is run on a collection of expert estimates that were made concurrently for a number of similar questions. In this case, we are aggregating estimates as opposed to temporal forecasts. These two different tests on real data provide an assessment and demonstrate the effectiveness of the two new methods proposed in this dissertation.

CHAPTER 2

FORECAST COMBINATION PUZZLE LITERATURE REVIEW

An understanding of the Forecast Combination Puzzle provides context around frequently used approaches to forecast combination and the preference for using fixed weights, a paradigm our approach improves upon.

2.1 Definition of the Forecast Combination Puzzle

The Forecast Combination Puzzle is the observation that when combining forecasts, equal weights (a simple average) often perform better out of sample than weights estimated to minimize the error variance in a sample of historic forecasts. Consider a group of k forecasts each with an error of e_i from the scalar true value θ. We will assume that the errors are unbiased and have a kxk positive definite covariance matrix (Σ). The forecast weights, $\vec{w}$, which minimize the squared error of the combined forecast, f_c, are (Bates and Granger, 1969; Winkler, 1981):

$$
\begin{aligned}
&\text{Let } f_i = \text{ forecast for expert } i \in (1 \dots k) \\
&\quad e_i = f_i - \theta \text{ where } \theta \text{ is the true value being forecast} \\
&\quad \vec{f} = (f_1 \dots f_k) \\
&\quad \vec{e} = (e_1 \dots e_k) \\
&\quad \vec{f} = \vec{1}\theta + \vec{e} \\
&\quad f_c = \vec{w} * \vec{f} \\
&\vec{w}_{optimal} = \frac{\vec{1}^t \Sigma^{-1}}{\vec{1}^t \Sigma^{-1} \vec{1}} \qquad (2.1) \\
&\text{where: } \vec{1} \text{ is a vector of } k \text{ ones}
\end{aligned}
$$

The variance and covariance parameters in Equations 2.1 and 2.2 are typically estimated from historical data by substituting parameter estimates for the true values. Using the notation from equation 2.1 the parameters can be estimated as follows assuming that the forecast errors are unbiased:

$$
\begin{aligned}
S^2 &= \vec{e}(\vec{e})^t \\
\hat{\mathbf{\Sigma}} &= \frac{S^2}{k-1} \\
\vec{\hat{\sigma}^2} &= diagonal(\hat{\mathbf{\Sigma}})
\end{aligned}
$$

We will define weights estimated using equation 2.1 as covariance weights. If the forecast errors e_i are normally distributed, this is also the maximum likelihood estimator. In all cases, the weights from Equation 2.1 are constrained to add to 1, however individual weights can be negative or positive. The model in Equation 2.1 shows that the optimal forecast weights is a function of the covariance matrix (Σ).

The accuracy and variability for these weights will then depend on the accuracy and variability of the estimates of the covariance matrix. The covariance model requires the estimation of $(k^2 + k)/2$ variance-covariance parameters to derive the k weights. Frequently in a forecast combination context, the number of forecasters will be larger than the number of historical data points available for each forecaster. This situation can result in more parameters for estimation than data points. In this paper we will call these the covariance weights.

If the forecast errors in Equations 2.1 are truly independent (or for convenience, assumed to be independent) then the covariance matrix becomes a diagonal matrix of error variances, σ_i^2, and Equation 2.1 simplifies to Equation 2.2. The optimal weights are determined only by the inverse of each expert's variance and only requires the estimation of k variance parameters to derive the k weights. In this paper we will call these variance weights.

$$w_i = \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}} \tag{2.2}$$

It is helpful to think of the inverse variance terms in Equation 2.2 as the skill of the respective expert. An expert with relatively small errors will have a smaller variance and a higher skill than an expert with relatively larger errors. We will assume that expert forecasts are believed to be unbiased and then estimate an expert's skill as the inverse of the mean squared forecast errors (MSE) adjusted for the variance sample size bias factor $n/(n-1)$.

Definition 2.1.1. *An expert's skill is the inverse of his error variance* $\frac{1}{\sigma_i^2}$.

Then for a two-forecast system, the optimal weight in Equation 2.2 can be expressed as the ratio of the two expert's skills (Eq. 2.3).

Definition 2.1.2. *Skill ratio (Skr), in a two-expert combination is the ratio of the two experts' skill.*

$$Skr_1 = \frac{\sigma_1^{-2}}{\sigma_2^{-2}}$$

$$w_1 = \frac{Skr_1}{1 + Skr_1} \tag{2.3}$$

$$w_2 = \frac{1}{1 + Skr_1} \tag{2.4}$$

As an expert's skill increases, the weights assigned to their forecasts will increase but at a slower rate. The weights will always be assigned in the same ratio as that of the expert's skill ratio. Although the most accurate forecast gets the greatest weight, all forecasts get some weight in this model as they have additional information based on the assumption of independent errors. Similar to covariance weights, variance weights are constrained to add to 1, but in this case they will always be bounded on the 0 to 1 interval as skill levels (inverse variances) can never be negative. We will refer to an expert's skill ratio throughout this paper as a way to consider his skill relative to the other experts in the forecast combination. In Section 4.2 we will define a skill ratio for one expert to multiple experts in a combination which will extend this concept to multi-expert combinations.

Empirical and simulation studies going back to Bates and Granger (1969) have frequently found that combinations which only use the forecast variances perform better than weights which include the estimated covariance terms (Bunn, 1985; Gunter, 1992; Newbold and Granger, 1974; Schmittlein et al., 1990; Smith and Wallis, 2009; Stock and Watson, 2004; Winkler and Makridakis, 1983). This result has been attributed to the errors in estimating the covariance terms and to the instability of the estimated weights as the cross forecast correlations increase, resulting in negative

weights in some cases (Aksu and Gunter, 1992; Claeskens et al., 2016; Gunter, 1992; Winkler and Clemen, 1992). The model proposed by Blanc and Setzer (2016) suggests that covariance optimized weights will always require a larger sample to estimate than variance optimized weights. In addition, Smith and Wallis (2009) found that the variance of covariance based weights will always be greater than that of variance based weights. Based on these observations, in this paper we will first focus on variance optimized weights in Chapters 3 - 4 but will extend the analysis to covariance weights in Chapter 5.

In the event that all the forecasts have the same variance, indicating that all forecasters are equally skillful, Equation 2.1 further simplifies to the weights all being equal and proportional to the inverse of the number of forecasts (a simple average):

$$w_i = \frac{1}{k} \tag{2.5}$$

The performance of equal weights is the "puzzler" as it often dominates its more complex brothers when the optimization of historical forecast errors suggests it should be otherwise. In this paper we will call this approach equal weights as each forecast gets the same weight. In this case, estimation of parameters is not required. The use of equal weights is a second, more commonly used, decision rule for the combination of forecasts.

2.2 Forecast Combination Puzzle Explanations

It is widely believed that the the puzzle can be explained by estimation errors in the weights offsetting the gain from using theoretically optimum weights (Bunn, 1985; Chan and Pauwels, 2018; Clemen, 1986; de Menezes and Bunn, 1998; Palm

and Zellner, 1992; Schmittlein et al., 1990; Smith and Wallis, 2009). The weight estimation error will depend on the distribution of the weights estimated from the historical performance sample. If the distribution is wide (high variance) due to noise or uncertainty in the forecasts, then the likelihood of estimating a less than optimal weight will increase. Blanc and Setzer (2016) analyzed the trade off between estimation error and prediction error using a distribution for the estimated weights in a two forecast combination suggested by Winkler and Clemen (1992). They developed a model for the sample size where the in-sample weight estimation error is expected to be the same size as the out-of-sample forecast combination prediction benefit from using covariance optimized weights. They found that for variance ratios near 1.0, a large sample size was needed, but this would decrease as the variance ratio increased. Schmittlein et al. (1990) used a simulation approach to develop thresholds for when more complex weight estimation methods are preferred over equal weights. They found a similar relationship between sample size and forecast variance ratios.

An additional explanation for the puzzle includes the sensitivity of the weight estimation to the changes in the forecast process or environment. Changes in forecasting methods or unexpected changes in conditions for the forecast can make the historical forecast performance sample less representative. Several studies have shown that equal weights are more robust to out-of-sample changes than estimated weights (Blanc and Setzer, 2016; Genre et al., 2013). These results suggest that, all things being equal, a decision maker should more conservatively use equal weights. These results also caution against the use of very large historical sample sizes to estimate future expert performance. Finally, Elliott (2011) has shown that equal weights and covariance optimized weights can mathematically be equivalent for specific covariance scenarios. In this case there is no puzzle, as the optimal weights are the same as equal

weights.

A more formal way to understand the puzzle is to look at the statistical properties of a combined forecast with known fixed weights (not necessarily equal) and one with estimated weights where the weights are considered an additional random variable with an associated probability distribution (Claeskens et al., 2016). Consider two unbiased forecasts of a scalar quantity θ each with unbiased errors e_i and error variances σ_i^2. We will further assume that the errors of the two forecasts are correlated to some degree ρ. In this case the weights w_i are fixed with no variability and are assumed to add to one. Then the expected value and variance of the combined forecast f_c become:

$$f_i = \theta + e_i \text{ for } i \in (1, 2) \text{ forecasts}$$

$$f_c = w_1 f_1 + w_2 f_2$$

$$\mathbf{E}[f_c] = \theta \tag{2.6}$$

$$Var(f_c) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2 \tag{2.7}$$

In this case the expected value of the combined forecast is unbiased (Eq. 2.6) and the variance of the combined forecast is only dependent on the variance of the component forecast errors and their correlation.

Equations 2.6 and 2.7 assume that the weights for the two experts are known constants. In practice, weights are estimated ($\hat{w}_i$) and have a weight estimation error (ϵ_i). The expected value and variance of the combined forecast (Equations 2.8 and 2.9) can be restated using estimated weights and their error terms (Claeskens et al., 2016) (see Appendix A for details). The interaction of the forecast error and weight estimation error terms become important to the combined forecast estimates. For

simplicity, both weight 1 and weight 2 are shown independently when in a two-forecast system they are directly related, if not equivalent.

$$
\begin{aligned}
\text{Let}: w_j &= \hat{w}_j + \epsilon_j \text{ for } i \in (1,2) \text{ forecasts} \\
f_c &= \theta + \hat{w}_1 e_1 + \hat{w}_2 e_2 + \epsilon_1 e_1 + \epsilon_2 e_2 \\
\mathbf{E}[f_c] &= \theta + Cov(\epsilon_1, e_1) + Cov(\epsilon_2, e_2) \qquad (2.8) \\
Var(f_c) &= \hat{w}_1^2 \sigma_1^2 + \hat{w}_2^2 \sigma_2^2 + 2\hat{w}_1 \hat{w}_2 \rho \sigma_1 \sigma_2 + \\
&\quad \sigma_1^2(\hat{w}_1^2 + \sigma_{w1}^2) + Cov(\sigma_{w1}^2, \sigma_1^2) - Cov(\epsilon_1, e_1)^2 + \\
&\quad \sigma_2^2(\hat{w}_2^2 + \sigma_{w2}^2) + Cov(\sigma_{w2}^2, \sigma_2^2) - Cov(\epsilon_2, e_2)^2 + \\
&\quad 2\hat{w}_1(Cov(e_1, \epsilon_1 e_1) + Cov(e_1, \epsilon_2 e_2)) + \\
&\quad 2\hat{w}_2(Cov(e_2, \epsilon_1 e_1) + Cov(e_2, \epsilon_2 e_2)) + \\
&\quad 2Cov(\epsilon_1 e_1, \epsilon_2 e_2) \qquad (2.9)
\end{aligned}
$$

Equation 2.8, shows that in the case of estimated weights, the forecast combination will be biased to the extent that there is covariance between the forecast errors and the weight estimation errors (Claeskens et al., 2016). Given that the weights are estimated from the forecast errors, it is likely there will be some covariance. This can be true even if the forecast errors are themselves unbiased. This bias term represents the in-sample weight estimation penalty and can be larger than the out-of-sample forecast combination error arising from using fixed weights.

In addition, the variance of the combined forecast is also impacted by the variance of the estimated weights and the interaction of the estimated weight errors with the forecast errors as described in Equation 2.9. The direction of the impact, will depend on the direction and level of the various covariances. This result suggests that estimated weights may not result in the lowest combined forecast variance based

on the shape of the estimated weight error distributions. Asymmetries in the weight estimation distribution will increase the overall forecast combination bias and variance through the forecast and weight error interaction terms in Equations 2.8 and 2.9 (Claeskens et al., 2016). The multiple expert analysis of the variances and expected values of a forecast combination have also been derived by Claeskens et al. (2016).

Taken together, the observations around the threshold for using estimated weights by Blanc and the theoretical analysis by Claeskens shows that there are rational explanations for the forecast combination puzzle. These explanations are both dependent on the shape of the estimated weight error distribution. The importance of the weight estimation error to the forecast combination supports the focus of this dissertation on reducing weight estimation errors.

CHAPTER 3

WEIGHT DISTRIBUTIONS AND CRITICAL SKILL RATIOS

In the previous chapter we discussed the strong performance of equal weight combinations and some of the theoretical explanations for this observed phenomenon. A key insight from Claeskens et al. (2016) is that the distribution of the estimated weight's error interacts through several covariance terms with each expert's error distribution to create biases and increased forecast combination error variance. Then assuming normally distributed expert errors, a skewed distribution of weight estimation errors will increase the forecast combination error variance through these covariances. However, Claeskens did not take the next step to evaluate the shape of the weight distributions. In this chapter we will better understand the shape of the weight distribution through simulations and then use this information to develop a new heuristic to distinguish experts who have forecast performance significantly different than the average of the crowd.

3.1 Distribution of estimated weights

Two different distributions have been proposed in the literature for the distribution of estimated weights. Dickinson (1973) derived a closed form probability density function for the distribution of one variance based weight (eq 2.2) in a two forecast combination (see derivation in Appendix B.2). Winkler and Clemen (1992) recognized that the estimation of a weight (eq 2.1) in a two expert system is the same as the following regression:

$$(\theta - f_2) = \hat{w}_1(f_1 - f_2) + e$$

The sample distribution of an estimated regression coefficient in a binormal distribution with correlation is known to be a Pearson VII distribution (Kendall, 1962). The Pearson VII distribution provides a closed form equation for the weight with expert correlation, while the Dickinson distribution assumes the two experts are independent.

As previously discussed in Section 2.1, variance based weights are just covariance weights with zero correlation. We can compare the Dickinson distribution to the Pearson VII distribution with a zero correlation term. In addition, a simulation of the distribution via bootstrap sampling (section 3.1.1) provides a third point of view on which distribution most closely approximates the true distribution. In all three cases it is assumed that a decision maker will estimate the skill (inverse variance) of each expert based on n historical forecast error values. We will assume that the errors are independent, unbiased, and normally distributed.

The distribution was also simulated for a sample size of n and expert error variance ratio of $R = \frac{\sigma_1^2}{\sigma_2^2}$ by taking n samples from a Normal(0,R) distribution for expert's 1 historical performance sample and another n samples from a Normal(0,1) distribution for expert 2's historical performance sample. These samples were used to estimate the variance of each expert's forecast errors and in turn calculate an estimated weight using Equation 2.2. This process was repeated 100,000 times to develop a distribution for the estimated weights. A comparison of the three estimates of the distribution are provided in Figure 1 for a sample size of 10 and a variance ratio of 4.

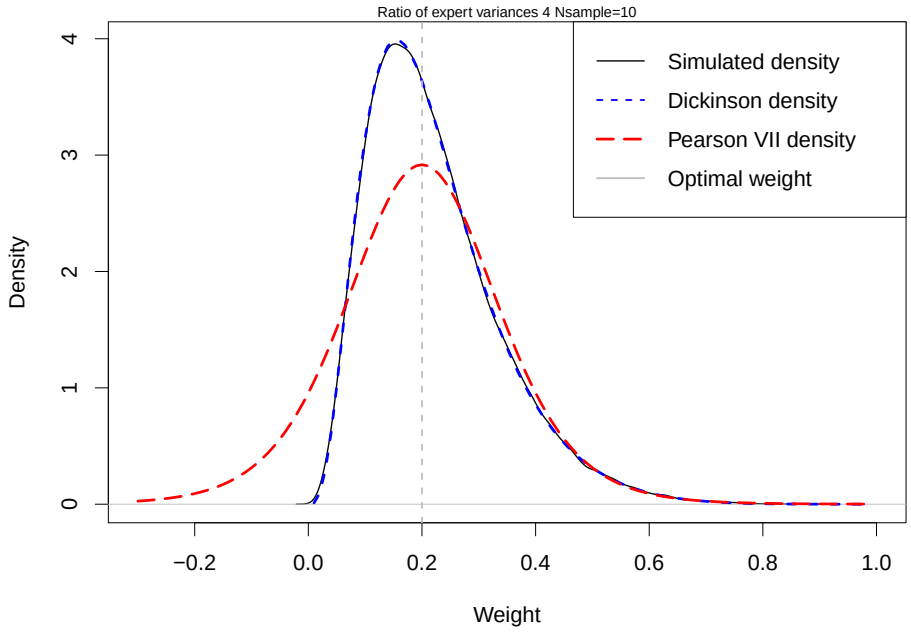
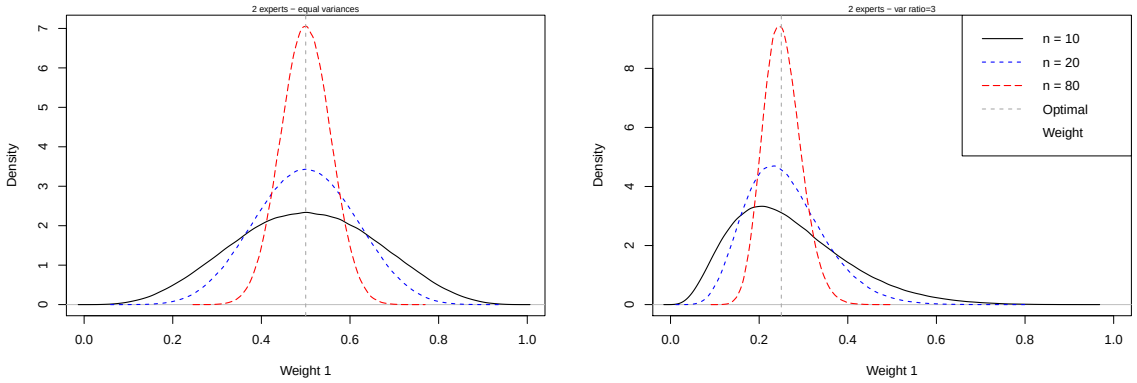


Fig. 1.: Comparison of predicted weight distributions to simulated values for a 2 expert system with 10 historical estimation points and a true expert variance ratio of 4 : 1

In this specific scenario, the Dickinson distribution in Equation 3.1 exactly follows the simulated density, while the Pearson VII distribution is substantially different. The Pearson VII distribution is centered on the optimum value of 0.2 for variance based weights while the Dickinson distribution has a pronounced skewness. The Pearson VII distribution assumes symmetry and is unbounded as it represents the distribution of a regression coefficient, however variance based weights are bounded on 0 to 1 as discussed in Section 2.1. The Dickinson distribution is bounded and therefore more closely follows the simulated weight distribution. All three distributions approach each other as the sample size increases and as the expert variance ratios approach one. Based on these observations, the Dickinson distribution is believed to be

the more representative distribution for variance based weights. The closeness of the Dickinson distribution to the simulation also confirms the simulation as an effective tool. We will use this simulation approach to generate estimated weight distributions for more complex conditions where a closed form solution may not be available.

The Dickinson density distribution (Eq. 3.1) for one forecast's estimated weight w_1 is illustrated in Figure 2 for various sample sizes (10,20,40) and expert skill ratios (1:1, 1:3). In all cases, the smaller sample sizes increase the width of the distribution significantly as there is less data to confidently estimate the weight. In the case of equal expert skill levels, the probability distribution function is symmetrical. However, whenever the expert skill levels are different, the density distribution will be slightly skewed at lower sample sizes. This skewness in turn will create a bias error in the combined forecast using estimated weights per Claeskens completing the explanation of the Forecast Combination Puzzle.



(a) Equally skilled experts (b) Expert 2's skill 3X expert 1

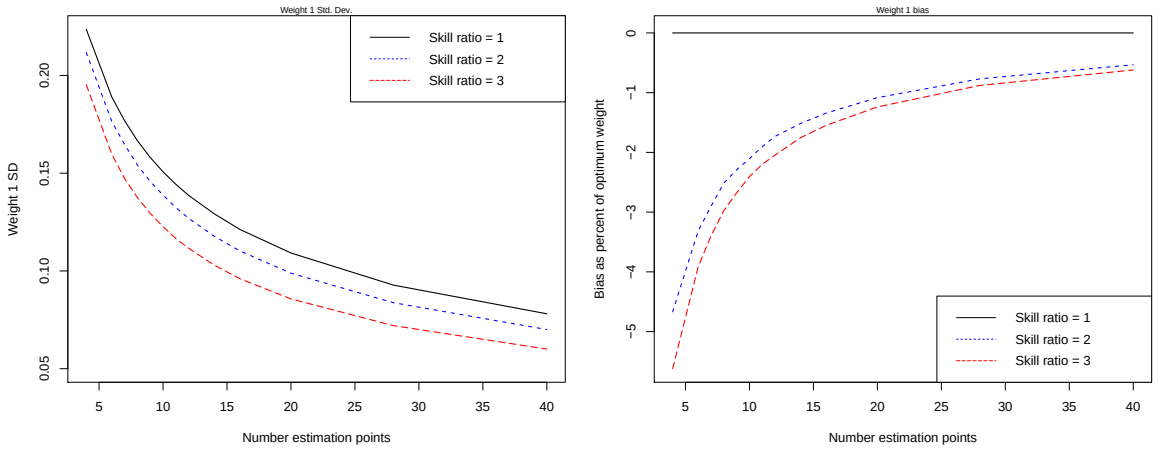
Fig. 2.: Impact of sample size and variance ratio on expert 1's estimated weight probability density in a two-forecast combination using simulated results.

The impact of the expert's skill ratio and estimation sample size on the standard

deviation and bias of the estimated weight can be seen in Figure 3. The standard deviation of an estimated weight is quite large given the $0-1$ range of weight values for sample sizes less than 25. The standard deviation of the estimated weight decreases as the number of estimation points increases. However the standard deviations also decrease with an increasing skill ratio for any given sample size. In this case, as the difference in skill becomes more apparent, the decision maker can be more confident in the differentiated weights. The percent bias in the estimated weight relative to the variance optimal weight decreases as the number of estimation points increase; however the rate of decline tapers off and the bias does not appear to approach zero for any reasonable number of sample points. This suggests that bias may always be a concern even in larger samples. The amount of bias increases with increasing skill level. At a skill ratio of 1.0, ie. equally skilled experts, there is no bias and the weight distribution is symmetric. As the difference in expert skill increases, the shape of the estimated weight distribution becomes increasingly skewed. However, most of the skewness appears as the skill ratio increases from 1 to 2. It does not appear to worsen going from a skill ratio of 2 to 3. In all cases, the bias will place less weight than is variance optimal on the higher skilled expert. The overall interaction of skill ratio on the estimated weight is complex as with increasing skill ratio, the variance of the estimated weight goes down providing a better estimate. But the estimate also becomes increasingly biased, reducing the accuracy of the estimate.

3.1.1 Boot strap approach to estimating weight distribution

Although there are proposed distributions for estimated skill based expert weights in two expert systems the author is unaware of a closed form solution for the distribution of estimated skill based weights in a multi-expert system. We will use a bootstrap approach to estimate the distribution of observed skill based weights in a multi-expert



(a) Standard deviation of an estimated weight

(b) Bias of an estimated weight

Fig. 3.: Impact of sample size and variance ratio on estimated weight distribution of one forecast in a two forecast ensemble using Equation 3.1.

system with various levels of true expert skills, number of experts, and true expert error correlations.

We assume that one expert has a specific true skill ratio (Skr_1) and all other experts (the crowd) have a true skill ratio of 1 and a true variance of 1. We further assume a true expert error correlation matrix $\mathbf{A}$ which is the identity matrix in the case of no expert correlations. In later sections we will assume an inter-class correlation matrix to model the impacts of expert correlation.

$$\mathbf{S} = \begin{bmatrix} \frac{1}{\sqrt{Skr_1}} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

$$\mathbf{\Sigma}_{true} = \mathbf{SAS}$$

$$\text{Sample:} \vec{X} = MVN(\vec{0}, \mathbf{\Sigma}_{true}) \ N \text{ times}$$

$$\mathbf{X} = (\vec{X}_1, \dots, \vec{X}_N)$$

$$\mathbf{\Sigma}_{observed} = \frac{\mathbf{X}\mathbf{X}^T}{N-1}$$

For variance based weights which assume no expert correlations we use the variances on the diagonal of the estimated covariance matrix to determine the weight of expert 1 per equation 2.2. The above process results in one set of estimated weights. This process is then repeated $100,000$ times to develop an estimate of the estimated weight distribution.

3.2 Confidence levels for using estimated weights

With a better understanding of the distribution for an estimated weight, a decision maker can assess the degree of risk that his estimate may not be optimal or not even better than an average. If the distribution is broad and/or biased, the likelihood that his estimate will be near optimal will decrease. Alternatively if the distribution is narrow and centered on the optimal value, the likelihood that the estimate will be near the optimal will be better. Unfortunately, based on the analysis in Section 3.1, we cannot expect the estimate weight distributions for small sample sizes found in forecast combination applications to be similar to the distribution of an estimated regression coefficient. This difference in distributions will make more traditional methods of confidence estimation less effective. We will revisit more traditional approaches in Chapter 4 when we consider multiple experts. For this chapter and future chapters we will simulate the distribution using an approach outlined in Section 3.1.1.

Traditional approaches to parameter estimation often ask if a parameter is different from 0. Given the objective of finding an estimation approach that is better than a simple average, we will ask whether the estimated weight is closer to optimal than an equal weight. By combining this revised question with a better understanding of the estimated weight distribution, we can develop an improved method for decision makers to decide on when to use estimated weights. Figure 4 illustrates this approach in a two-expert system.

Consider a two-expert combination where the first expert is twice as skillful at forecasting the outcome as the second. If we know that this forecaster is twice as skillful then this forecast deserves more weight. The optimal variance based weight for expert 1 is 0.67 per Equation 2.2 while the equal weight is only 0.50. However,

the only evidence that the decision maker has of the expert's relative skill levels is a historical sample of n past forecasts and forecast errors from each expert. There is a risk due to natural variation in the forecast errors that these samples may not fully represent expert's 1's skill and the decision maker could over or under estimate the skill of expert 1 relative to expert 2. Figure 4 provides the distribution for the weights that the decision maker could estimate depending on the errors in the historical performance sample.

Figure 4a shows the distribution of the weight estimated from a sample of ten previous forecast errors. In this case the decision maker could estimate a weight anywhere from ~ 0.2 to ~ 0.9 depending on the quality of the historical sample when the optimal weight is 0.67. An equal weight for this combination of two experts is 0.5. If the decision maker estimates the weight as 0.6 then this estimated weight will be closer to the optimal weight and this estimate will perform better than the equal weight of 0.5. However, the region in blue shows that there is a significant risk that he may estimate a weight that is less than 0.5, in which case the equal weight would be a better choice. The decision maker also has a risk of estimating a weight that is too high. The distance of an equal weight from the optimal weight is $0.67 - 0.5 = 0.17$. If the decision maker were to estimate a weight on the high side that was a greater distance from an optimal weight than an equal weight, the equal weight would be a better choice. This reflection of the distance from optimal assumes that the decision maker's loss function is symmetric which would be true for absolute or squared errors. In this case, any estimate of weights greater than $0.67+0.17 = 0.84$ would result in an estimate that is further away from optimal than an equal weight is. In summary, the areas in blue show the likelihood that an equal weight will be better than an estimated weight for the given estimated weight distribution. This leads us to a definition for a decision maker's confidence in using an estimated weight:

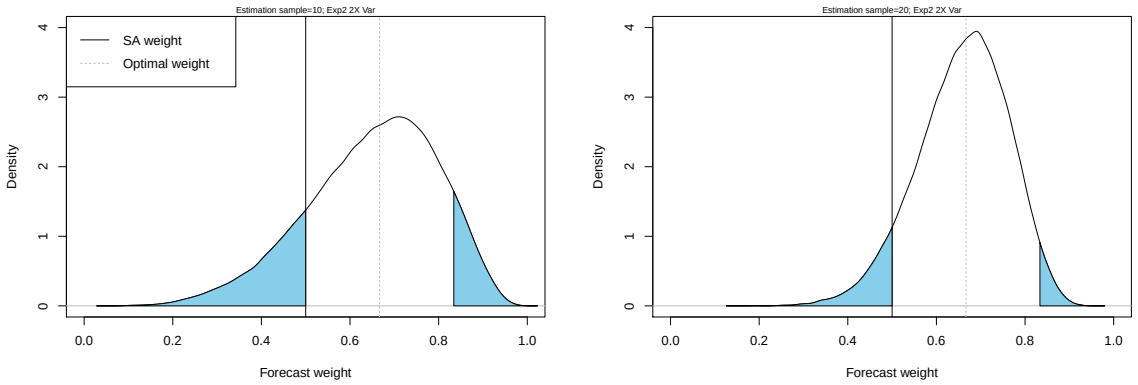
Definition 3.2.1. *A decision maker's confidence level that an estimated weight will be better than an equal weight is the area under the weight distribution curve between the optimal value of the weight and +/- delta, where the delta is the absolute value of the optimal weight minus an equal weight.*

We can see from Figure 4B where the estimation sample size is 20, the decision maker's confidence that an estimated weight is much higher, as he has a much better idea of the true relative performance of the experts. In general, the decision maker's confidence level will be determined by the sample size and the relative skill ratio of the expert relative to the other experts. This definition of confidence applies to multiple experts and scenarios with expert correlation provided the weight distribution curve is estimated accordingly. Both the concept of associating a confidence level with making the estimation decision and this approach to estimating a confidence level are new contributions to the field of forecast aggregation. We will expand the use of this approach to multiple experts in Chapter 4 and to scenarios with expert correlation in Chapter 5.

3.3 Critical skill ratio for using estimated weights

The confidence level to use an estimated weight assumes that the true skill level of the experts is known in order to develop the weight distribution curve. In practice, the decision maker will only know his available sample size and desired level of confidence. From this information, a skill ratio and associated weight distribution curve can be found where the confidence that an estimated weight equals the decision maker's desired level of confidence given the available sample. This leads to a second definition:

Definition 3.3.1. *Critical skill ratio is the expert skill ratio required such that the*



(a) 10 estimation points (b) 20 estimation points

Fig. 4.: Region where SA is a better weight estimate than Var based weights in a 2 forecast combination where expert 1 is twice as skillful as expert 2. The areas shaded in blue represent the probability that a simple average will be a more accurate than an estimated weight.

confidence level of using an estimated weight equals the decision maker's desired level of confidence given the available estimation sample.

If the observed skill ratio of an expert relative to other experts in the combination based on the historical performance sample is closer to 1.0 than the critical skill ratio then the expert is not sufficiently differentiated from his peers to meet the decision maker's level of confidence for using a differentiated weight. In this case, an equal weight should be used. However if the critical skill level is further from 1.0 than the critical skill ratio then it is likely that the decision maker's level of confidence will be exceeded and an estimated weight should be used. Note that this is still a probabilistic decision and there will remain a non-zero probability that this could have a less desirable outcome. However, in this approach the decision maker knows the level of risk he is taking in making the decision. This approach to defining critical

skill ratios is an improvement on the work of Blanc and Setzer (2016) in that the decision maker can select his desired level of confidence and the model is extensible to multiple experts with expert correlations.

There are two different situations and types of critical skill ratios to consider, sufficiently higher skill, and sufficiently lowered skill. An expert could have a higher skill than his peer(s) in which case the observed skill ratio should be above the critical skill ratio to justify an estimated high weight. Alternatively, an expert could have a lower skill than his peer(s) and if it is lower than the critical low ratio an estimated low weight is justified. In the case of a two model combination these respective high and low critical ratios should be reciprocals but this may not be the case in larger combinations.

3.3.1 Critical skill ratio estimation

A numerical approach to estimating the high and low critical skill ratios for a given level of confidence and number of available estimation points has been developed using the bootstrap skill based weight distribution from section 3.1.1. A numerical approach is preferred as it will be extensible to multiple expert combinations and to including expert correlations. We create a subroutine that estimates the confidence that a skill based weight will be better than a simple average for a given level of true expert skill and correlations. Then conduct a search to find the skill ratio where the confidence that the observed skill based weight performs better than a simple average equals the desired level of confidence.

3.3.1.1 Confidence estimation

We can estimate the confidence level that a skill based weight will perform better than a simple average by determining the areas under the skill based weight distribution curve that are further from optimal than a simple average (Figure 4). The procedure used is as follows:

1. Estimate the weight distribution for the given skill ratio (Skr_1), number of experts N, the true expert correlation structure, and for the assumed weight estimation method (variance, covariance, or any other approach). Note in this section we are assuming no expert correlation so the correlation structure is an identity matrix. Also the given skill ratio will be varied as part of the golden search process until the desired confidence level is achieved.

2. Calculate a histogram of the weight distribution using 400 breaks, calculate the minimum estimated weight W_{min} and the maximum W_{max} in the distribution.

3. Calculate the optimal skilled based weight W_{opt} for the given skill ratio using the desired skilled based weight method.

4. Calculate the simple average weight $W_{avg} = \frac{1}{N}$ for the number of experts under consideration.

5. If $W_{avg} <= W_{opt}$ then use the histogram to estimate the area under the weight distribution from W_{min} to W_{avg} (left tail) and the area from $2W_{opt} - W_{avg}$ to W_{max} (right tail). We estimated the area by summing the densities of each histogram bar in the region of interest.

6. If $W_{avg} >= W_{opt}$ then use the histogram to estimate the area under the weight distribution from W_{min} to $2W_{opt} - W_{avg}$ (left tail) and the area from W_{avg} to

W_{max} (right tail).

7. Calculate the overall confidence that an estimated skill based weight will perform better as 1 minus the percent of the area under the two tails from steps 5 or 6.

3.3.1.2 Critical skill ratio search

We can use the confidence estimation approach (Section 3.3.1.1) to guide a search for the skill ratio that causes the estimated confidence to equal the decision maker's desired confidence (See critical skill ratio definition 3.3.1). The procedure is described below. This overall approach can be repeated for various methods of skill based weight calculations, levels of confidence, number of experts, and correlation assumptions to create pre-calculated tables for a decision maker's use.

1. Estimate the confidence that a skill ratio of 1 and a skill ratio of 10 is better than a simple average per Section 3.3.1.1

2. Use the uni-root function in R to find the skill ratio between 1 and 10 where the difference between the estimated confidence and the desired confidence is zero within a desired tolerance (0.002). This is the high critical skill ratio.

3. Estimate the confidence that a skill ratio of 0.1 and a skill ratio of 1 is better than a simple average per Section 3.3.1.1

4. Use the uni-root function in R to find the skill ratio between 0.1 and 1 where the difference between the estimated confidence and the desired confidence is zero within a desired tolerance (0.002). This is the low critical skill ratio.

3.4 Critical skill ratio sensitivity

This approach in Section 3.3.1 has been used to estimate the relationship between critical skill ratio, number of sample points, and the desired confidence level is illustrated in Figure 5 for a two expert combination. As the number of historical sample points increases, the underlying estimated weight distribution narrows, and a smaller relative skill ratio can confidently be discerned. As the level of desired confidence increases, a larger skill ratio that is further from one is required for the same estimation sample size to provide a more confident decision.

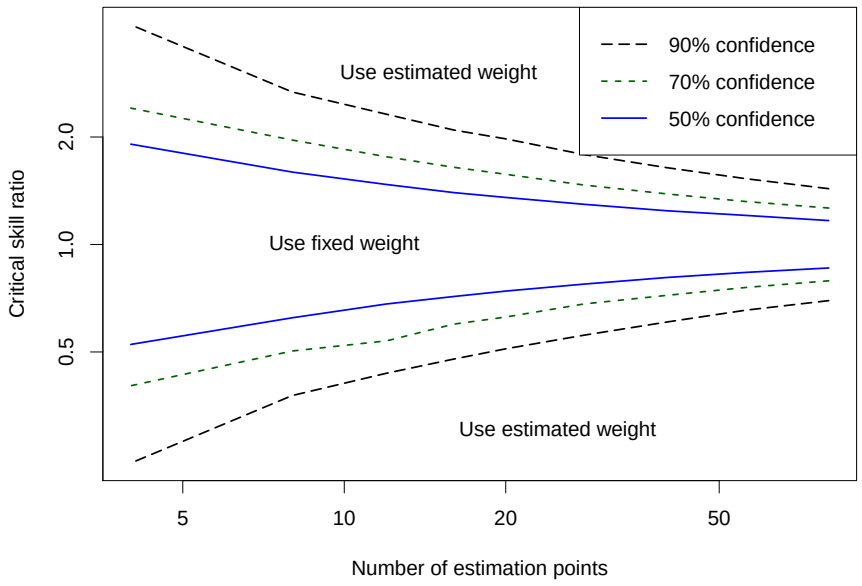


Fig. 5.: Region of skill ratios where fixed weights should be used at various levels of confidence based on the number of historical skill estimation points.

3.5 Critical skill ratio test

The critical skill ratio (CSKr), high or low, can be used as a test by a decision maker to determine if an estimated variance based weight should be used instead of an equal weight for each expert's forecast in his combination. The region where equal weights should be used is bounded by the two critical skill ratio curves as illustrated in Figure 5. When an expert skill ratio is above or below the critical skill ratio, estimated weights should be used. For a two-expert combination both experts will either be outside or inside the lines, this may not be the case when we consider multi-expert combinations.

The efficacy of this new decision heuristic was validated in a two expert simulation using an approach similar to that used by Schmittlein et al. (1990). Twenty-seven synthetic forecasters were created with three sets of nine expert skill levels ranging geometrically from 0.5 to 2.0. For each expert, 140 forecasts were randomly drawn from a normal distribution with no bias and a variance corresponding to the inverse skill ratio. The first 40 forecasts were used to generate estimation samples of sizes $4-40$ for randomly chosen sets of two experts. The remaining 100 points were used to evaluate the frequency that the proposed differentiation approach had a lower absolute error than using equal weights. 2000 expert forecast combinations were sampled for each estimation sample size of 4,8,12,16,20,28, and 40. The results are summarized in tables 1 and 2.

The forecast combinations flagged for differential treatment by the critical skill ratio decision at a 90% confidence in Table 2 achieved an 86% selection accuracy rate for an estimation sample size of 4. However, this accuracy decreased to below 80% with increasing estimation sample. This suggests that the critical skill ratios may

Estimation sample points	4	8	12	16	20	28	40
Frequency correct estimation decision	85.8	85.0	83.2	84.3	82.3	77.8	76.2
Frequency correct equal weights decision	50.6	49.1	47.7	48.0	48.3	48.6	48.5
Percent forecasts differentiated	26.9	31.3	37.3	36.5	42.8	51.0	54.2

Table 2.: Simulation results for a 2 expert combination with a critical skill ratio confidence level of 90%.

have been under estimated. Even so, the overall accuracy is close to the expected 90% level associated with a 90% confidence. It is interesting to note that the percentage of forecasts exceeding the critical skill ratio and receiving a differentiated weight nearly doubled as the estimation sample size increased. Then the fall-off in selection accuracy is more than compensated by the much higher number of forecasts accurately being chosen for a differential weight. Although the accuracy of this test to flag a forecast for differentiation is good, the power to resolve when it is more accurate to not differentiate is poor. The test was only $\sim 50\%$ accurate on the decision to not differentiate. This is not a surprise as the test was constructed to favor equal weights as a more conservative approach when there is not clear evidence of differentiated forecast skill. The rate at which forecasts were flagged for differentiation increased substantially as the estimated sample size increased. This is consistent with the region of no differentiation in Figure 6 shrinking as the estimation sample size increases. The impact of sample size on the region of no differentiation is most pronounced for the 90% confidence level curves.

The forecast combinations flagged for differential treatment at the 50% confidence level in Table 3 achieved a much higher relative accuracy in the high 60% range than the expected confidence level of 50%. This suggests that the critical skill ratios may have been over estimated. Similar to the 90% confidence simulation, we

Estimation sample points	4	8	12	16	20	28	40
Frequency correct estimation decision	67.7	67.5	68.3	67.1	67.1	64.4	63.7
Frequency correct equal weights decision	50.4	50.3	49.1	49.8	50.4	49.6	49.7
Percent forecasts differentiated	61.8	67.3	69.3	73.2	73.7	78.7	80.4

Table 3.: Simulation results for a 2 expert combination with a critical skill ratio confidence level of 50%.

see the same but much less pronounced trend of decreasing accuracy as the estimation sample size increases. Also, similar to the results at a 90% confidence level the power of the test to accurately determine when not to differentiate forecasts is low for the same reasons. The rate at which forecasts were flagged for differentiation also increased but with a much less pronounced increase which is consistent with the 50% critical skill ratio curves being less sensitive to sample size than the 90% curves.

Overall, the relatively high accuracy at each confidence level suggests that the critical skill ratios estimated are effective discriminators of when to use a differentiated weight in a two forecast ensemble. Even so, there may be some opportunities to improve the estimation of critical skill ratios and/or optimize the critical skill ratios to provide a decision maker with the highest probability of achieving a more accurate combination than using equal weights.

3.6 Summary

In this chapter we have developed the foundation for a new heuristic to determine when to use estimated weights. We have approached this problem in two ways that are new to the Decision Sciences literature on forecast aggregation:

1. **Distribution assumptions** - Past work on decision thresholds and the variability of weights have assumed a symmetric weight estimation distribution

(Blanc and Setzer, 2016; Winkler and Clemen, 1992). This is a reasonable assumption for larger sample sizes, but is unlikely for the sample sizes typically encountered in forecast combination. A simulation process has been developed to estimate the true weight distributions at low sample sizes which will improve how these decisions are made. The simulation process will also enable extensions to multi-expert combinations with expert correlations.

2. **Confidence levels for the estimation decision** - Past work has developed a decision threshold for estimating weights based on equal expected losses between using a simple average and estimated weights (Blanc and Setzer, 2016). A decision maker may expect more than equal benefits before deviating from the well accepted method of using a simple average. We have introduced and estimated the confidence level a decision maker can have in this decision. Considering the confidence level in a decision is a more informative way to frame the alternative estimation approaches.

Both of these improvements have been achieved without greatly increasing the complexity of the decision maker's process. Tables of critical skill ratios for various estimation sample sizes can be estimated once and made available. A decision maker only need to check where he is on the threshold table.

In Chapter 4 we will develop new and deeper insights on how multiple experts impact the weight estimation process. We will also extend the concept of critical skill ratios to multi-expert combinations. In Chapter 5 we will add expert correlations to the analysis. In this chapter we develop a second, new heuristic for forecast combination that is complementary to using critical skill ratios. In addition, we extend the critical skill ratio approach to more common scenarios where expert correlations are present. Finally, Chapter six provides a comparison of the two newly proposed ap-

proaches, as well as, several current benchmark approaches on two different empirical sets of forecast data.

CHAPTER 4

COMBINATION OF MULTIPLE FORECASTS

In this chapter we will extend the approach developed in Chapter three for 2 expert combinations to combinations of multiple experts. The dynamics of weight estimation in a mulit-expert combination will be explored and a revised critical skill ratio based on the principles previously discussed will be developed for multiple expert combinations. Two heuristics to improve on forecasts combinations using a multi-expert critical skill ratio will be proposed and validated with a simulation.

4.1 Literature review - Combination of multiple forecasts

Two approaches to the combination of multiple models have been explored: differential weighting and differential inclusion.

4.1.1 Differential weighting past work

The robustness of the simple average for the combination of multiple models versus using a skill based weight has been observed in several studies over the years. In an early study of time series forecasting techniques, Makridakis (1982) found that the simple average of six forecasting methods performed better than a weighted average using covariance based weights. In 2004 Stock and Watson (2004) reported on a comparison of 49 forecasting methods across 215 US macro economic time series and found that a simple average performed better than a weighted average. Clemen (1986) found similar results when combining forecasts from four econometric models. Genre et al. (2013) asked the question " Combining expert forecasts: Can anything beat

the simple average' after reviewing the aggregation of the European Central Bank Survey of Professional Forecasters data for six variables from 1999 - 2011. They found instances where some more complex aggregation schemes beat a Simple Average but could not identify an overall pattern of better performance and concluded that the consistent performance of a simple average was the better choice. In all of these cases, the forecast quantity was a time series.

Schmittlein et al. (1990) explored under what conditions would a more complex weighting scheme beat a simple average through a simulation study of various forecast variance ratios and levels of expert correlation. He proposed using the Akaike's Information Criteria (AIC), applied to the historical estimation sample, to select the most predictive model when combining more than two forecasts. In this case, the expert weights from each of the combination methods being assessed (SA, Var, Cov) are models. Then the model residuals would be the errors of the combined forecast evaluated at each of historical forecast data points in the estimation sample. One can assume normally distributed errors and calculate the sum of the log likelihood of the errors in the estimation sample for the weights associated with each method. The log likelihood is then penalized for the number of parameters being estimated which in this case will vary for each model. The AIC statistic can be calculated for each combination method per the following equation. The method with the lowest AIC is selected to perform the out of sample forecast combination.

$$AIC_{simple\ average} = -2\sum_{i=1}^{n} log(\pounds(e_i))$$

$$AIC_{variance\ weights} = 2k - 2\sum_{i=1}^{n} log(\pounds(e_i))$$

$$AIC_{covariance\ weights} = k(k-1) - 2\sum_{i=1}^{n} log(\pounds(e_i))$$

Where

$k =$ number of experts in the combination

$n =$ number of estimation data points

$e_i =$ the combined forecast error at data point i

This approach in a five forecast combination selected the most accurate model in 66% of the cases but did not outperform a simple average overall. The AIC approach over weighted the selection of the most complex, covariance model by a factor of 3. AIC is only one of many regression model selection techniques, other statistics include the Bayesian information criteria (BIC), corrected AIC (AICc), as well as others (Hurvich and Tsai, 1989; Schwarz, 1978). However, AIC is the only statistic that has been applied to forecast combinations that the author is aware of, as such we will use it as a benchmark.

The key question for differential weighting is when is it advantageous to use skill based weights instead of a simple average in mulit- expert combinations? Clearly in many cases a simple average has the most accurate performance but there are cases where skill based weights out perform a simple average. How should a decision maker decide when to use skill based weights? We will propose a heuristic based on a critical skill ratio and compare its performance to the AIC method.

4.1.2 Differential inclusion past work

Several studies have found that the improvement of a simple average combination over individual forecasts diminishes as the number of models increases, with minimal improvements observed after 5 models are included (Clemen and Winkler, 1985, 1999; Makridakis, 1982; Newbold and Granger, 1974). Using this principle, Mannes et al. (2014) investigated the performance of choosing the best expert, the average of a select crowd of N experts, and the average of the full crowd. They looked at how these combination methods performed in environments described by two dimensions:

- **Dispersion in expertise** The degree of individual forecast variation around the mean of all forecasts in the combination.
- **Level of bracketing** The degree that forecasts are on either side of the realized value (ie. some higher and some lower) as opposed to all being on the same side of the realization (ie. all higher or all lower)

In a simulation study they found that choosing the best (most accurate) expert strategy performed well in high dispersion environments with low bracketing while the average of all experts worked well in low dispersion environments with high bracketing, but that the selection of the best few experts performed well in mixed environments and was robust even in the more extreme environments. A subsequent application of the select crowd principle to 90 different data series found that a strategy of averaging the top 5 performers was the most accurate performing strategy in 61% of the data sets and was only the worse in 3% of the data sets.

An alternate approach to choosing the top 5 experts, is to exclude the outliers either through a trimmed mean or a winsorized mean (Jose and Winkler, 2008). In an extensive study of trimmed and winsorized means the authors found that when

the coefficient of variation for the forecasts in the ensemble was high, the trimming or winsorizing performed better than the overall mean. However for low coefficients of variation the overall ensemble mean was more accurate. An additional advantage of this trimming approach was the reduction in the frequency of large forecast errors. The authors recommended either trimming by 5-15% on each side or winsorizing by 7.5 -22.5% on each side.

Researchers have also suggested using a best subset approach where the most accurate simple average of every possible subset of expert combinations would be used to select the optimum group of forecasters (Bunn, 1981; Elliott, 2011). This method is attractive at face value due to its simplicity but an analysis of 20 or more experts as typically found in professional surveys would require $2e18$ different combinations to be compared which may not be practical.

In summary, it is clear that including all experts in a larger combination does not enhance performance in practice. Using a simple average of a few, best experts appears to be the most robust approach. A key question for differential inclusion is how to choose which experts and how many to include? A second question, is for the selected set of experts, is a simple average the most accurate approach or should skill based weights be used? We will propose a second heuristic for the selection of experts based on critical skill ratio and compare its performance to selecting the top 5 experts.

4.2 Impact of multiple forecasts on estimated weights

When considering the impacts of multiple experts in a combination, it is helpful to ask how do multiple experts impact the most outstanding expert in the crowd. Let expert 1 be the most skillful expert in a combination with a crowd of $2 \dots k$ other experts. Based on how expert variances enter the optimal weight Equation 2.2, the

average skill of the crowd will be the average of the inverses of each expert's variance (see Appendix C). A less often used average, the harmonic average, is the inverse of the average of the inverses of the values being averaged and is often used when averaging rates. This leads us to the definition of a multi-expert skill ratio:

Definition 4.2.1. *Multi-expert skill ratio is the ratio of the inverse variance of an expert to the inverse of the harmonic average of the variances of the other experts.*

$$Skr_1 = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{k-1}\sum_{j=2}^{k}\frac{1}{\sigma_j^2}} = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{HA(\sigma_{2\ldots k})^2)}}$$

Where HA() is the harmonic average.

A corollary to the definition of the Multi-expert skill ratio is that the variance of a crowd of experts is measured as the harmonic average of the variances of each crowd member:

$$\sigma^2_{crowd} = HA(\sigma^2_{1\ldots n}) = \frac{n}{\sum_{i=1}^{n}\sigma_i^{-2}}$$

The Harmonic average is one of three Pythagorean means. It has the property that it will always be less than the more conventional arithmetic mean for positive numbers such as variances (Maor, 1977). The author started by using an arithmetic mean to normalize data, but found that the harmonic mean provided a clearer picture of trends and relationships when working with combinations of multiple forecasts.

We can restate Equations 2.3 and 2.4 for a multi-expert combination in terms of the multi-expert skill ratio as follows (see Appendix C). As long as multi-expert skill ratio as defined is used, these Equations are universal for $2 \ldots k$ experts.

$$w_1 = \frac{Skr_1}{(Skr_1 - 1) + k} = \frac{Skr_1}{Skr_1 + (k - 1)} \tag{4.1}$$

$$w_{crowd} = \frac{1}{(Skr_1 - 1) + k} = \frac{1}{Skr_1 + (k - 1)} \tag{4.2}$$

From Equations 4.1 and 4.2 it can be seen that the weight of expert 1 will always be weighted more heavily than an individual from the crowd in proportion to his relative skill level regardless of crowd size. However, when an average skilled expert is added to the crowd the skill ratio will remain unchanged but overall weight given to the better expert (expert 1) will go down, essentially allowing an increasingly large, average skilled crowd to shout out the most accurate expert. This, in part, explains why using the best few experts works well. The crowd shouting effect is capped by limiting the combination to 5 experts.

The denominators for the estimated weights in Equations 4.1 and 4.2 will be dominated by k, the number of experts, when the number of experts k is much larger than the best expert's skill ratio. In these cases, k acts like a scale term and the estimated weight distribution will be scaled down and become more narrow as k increases. However, even if k becomes very large, the ratio of the weight given to expert 1 as compared to the average weight for the rest of the crowd will remain fixed at the expert's skill ratio. This suggests that the skewness observed in the two forecast combinations with unequal skills will persist to all sizes of expert combinations. Also, as the skewness persists and the distribution width goes down with increasing k, we will see the skewness increase on a percentage basis which will in turn increase the estimation error.

Alternatively, in the case of smaller collections of experts the number of experts can very well be of the same magnitude as the skill ratio of the best expert. In these

cases the denominator of the estimated weight distribution will be equally impacted by skill ratio and number of experts. We will see how the skill ratio interacts with the number of experts in the weight simulations in Section 4.2.1.

Equations 4.1 and 4.2 can be used to develop an estimate for the overall forecast combination variance (see Appendix C) as follows:

$$\sigma^2_{combination} = \frac{\sigma^2_{crowd}}{(Skr_1 - 1) + k} \tag{4.3}$$

In this case, the $(Skr_1 - 1) + k$ term acts like an effective sample size. If the skill ratio of expert 1 is better than average then the effective sample size of the combination is larger than the k members and the overall combination error is smaller and vice versa. If expert 1 has the same skill as the crowd ($Skr = 1$), then the sample size remains k ,the number of experts, and we have equal weights.

Another way to consider Equation 4.3 is to redefine the formerly outstanding expert 1 to being the last expert to have been included in the combination. Prior to adding this last expert, the forecast combination variance would have been:

$$\sigma^2_{combination} = \frac{\sigma^2_{crowd}}{k - 1} \tag{4.4}$$

In this case, the forecast combination variance with the last expert in Equation 4.3 will always be lower than the forecast combination variance before the last expert (eq 4.4) as the skill ratio can never be negative. This improvement in the overall forecast combination variance is true regardless of how poorly skilled the last expert is. This phenomena is a result of our independence assumption which we will explore in Chapter 5.

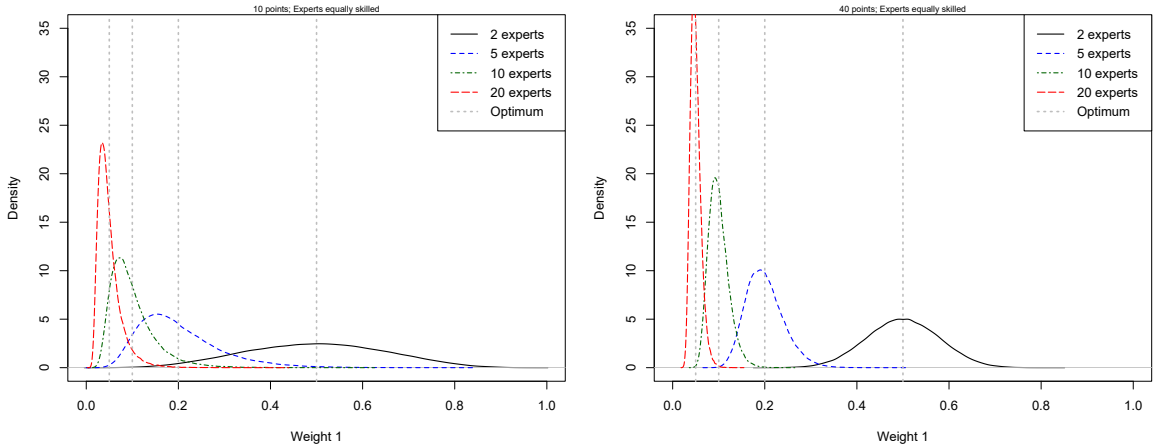
4.2.1 Impact of number of experts on weight distributions

In the two expert combination we have used the distribution of the weights to find a critical skill ratio for when differentiated weights should be used. We will follow a similar approach in the analysis of multiple expert combinations. The distribution of one expert in a multi-expert combination was simulated using the same approach outlined in Section 3.1 by sampling k experts instead of just 2 experts (Figure 6). In Figure 6A, with only 10 estimation points and with equally skilled experts, the width of the distribution for the estimated weight decreases as the number of experts increases from 2 to 20 experts. However, we also see that the distribution is skewed for expert combinations with more than 2 experts. The symmetry in the weight estimation process, even when equally skilled, is lost when a third or more expert is added.

The vertical gray dotted lines represents the variance optimal weight for the given skill ratio in Figure 6. In this case, the expert skills are all equal and the optimal value is equal weights. Then, as there are more experts, the optimal equal weights vertical line moves to the left along with the overall distribution. Except for the two expert case, the weight distributions are all skewed to lower values than is optimal. This skewness is induced by the likely presence of estimation error in at least one of the expert's skill. Even when the expert skills are equal, there is a likelihood that one of the many experts in the combination will be erroneously estimated to receive a higher weight. When this error occurs, all of the other weights will receive a lower skill than deserved because the weights are constrained to add to one. The end result is a skewed distribution. The impact of the number of experts provides a new dimension and insight to the forecast combination puzzle. Adding more experts, even when holding the skill ratio and estimation sample size constant, will increase the

skewness of the distribution, which in turn increases the estimation error and causes equal weights to be more favored.

The impact of an increased weight estimation sample size is illustrated in Figure 6B. As observed in the two-expert combination, increased estimation sample narrows the weight estimation distribution reflecting an increased level of confidence in the estimated weight. The skewness in the distributions persists but is at a lower level.



(a) 10 estimation points (b) 40 estimation points

Fig. 6.: Simulated estimated weight distribution for one expert in a multi-expert combination with equal skills.

The impact of increasing skill level on the weight distribution of an expert in a 10-expert combination is illustrated in Figure 7. Increasing the expert's skill ratio also increases the width of the distribution for the estimated weight. This is the opposite of what we observed in the two expert system (Figure 2, Figure 3a) but can be explained by Equation 4.1. With a large number of experts, the denominator of the estimated weight in Equation 4.1 becomes insensitive to changes in skill ratio. However the numerator in Equation 4.1 is only the skill ratio and increasing the skill

ratio will have the effect of scaling the distribution to higher and wider values. This is an important insight as we can not expect to extrapolate behaviors from two-expert combinations to multi-expert combinations as the distribution dynamics are different - in this case reversed.

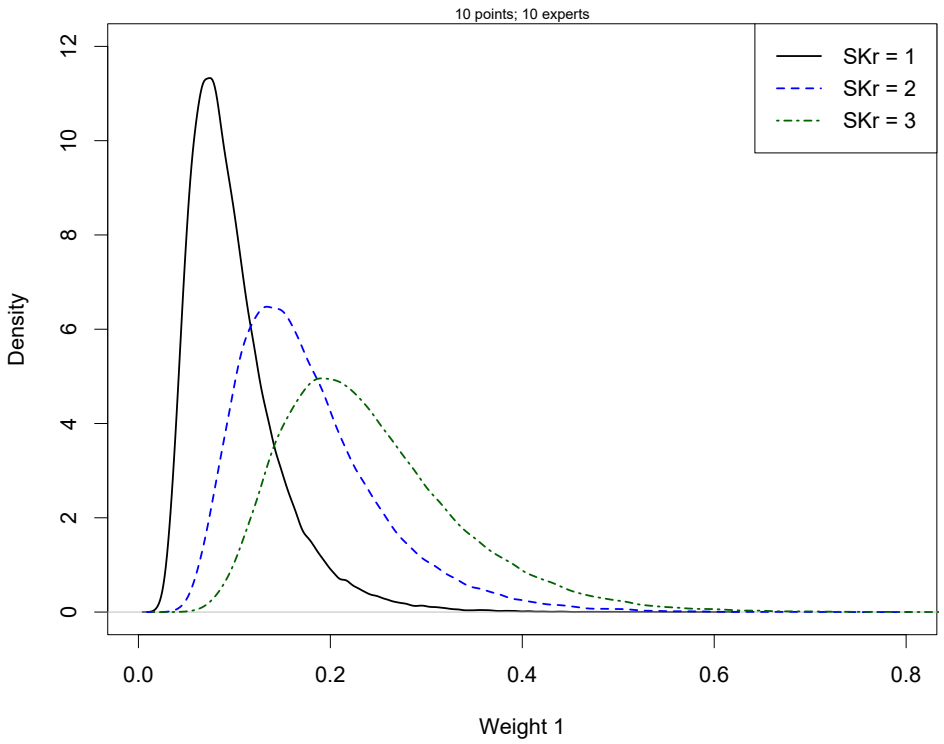


Fig. 7.: Simulated estimated weight distribution for one expert with various skill levels in a 10-expert combination with 10 historical estimation points.

Changes in the bias of the simulated distributions for an estimated weight with multiple experts are illustrated in Figure 8. In this case, we define bias as the median of the estimated weight distribution minus the variance optimal weight for the given skill ratio. As the bias increases, the likelihood the decision maker will estimate a weight that is different from the optimal value will increase. The bias is induced by the skewness in the distribution caused by the $0 - 1$ constraint on the estimated weight as seen in Figures 6 and 7. For the bias in smaller combinations of experts,

there is a clear interaction between the skill ratio and the number of experts. The bias increases with skill ratio until the number of experts is sufficient to diminish the bias. This interaction can be explained by the denominator of the estimated weights in Equations 4.1 and 4.2. When $k-1$, the degrees of freedom due to the number of experts becomes large relative to the skill ratio, the number of experts will begin to dilute the bias effect of skill ratio and the bias decreases. Even so, the bias continues even for larger groups of experts and in general, the more highly skilled expert will have a greater bias error consistent with the scaling effect that skill ratio has on the weight of expert 1 in Equation 4.1. In the case of a larger estimation sample size (dashed lines Figure 8), the same pattern is present but to a lower degree due to the increased accuracy of the larger sample size estimates. In all cases the bias under weights the more skillful expert. Overall, these observations support the forecast combination puzzle by demonstrating the difficulty of obtaining an unbiased estimate of the optimal weight.

Changes in the standard deviation of the simulated distributions for an estimated weight are illustrated in Figure 9. The general trend of decreasing standard deviation for the estimated weight with an increasing number of experts is seen as expected only when the number of experts is higher than the skill ratio of the expert. This is the wisdom of the crowd effect where a larger sample of experts will tend to average-out the noise. In both Equations 4.1 and 4.2 when the number of experts increases the overall weight decreases. In addition, the use of more estimation sample points (dashed lines) also decreases the standard deviation in all cases similar to a two-expert combination. The larger sample size used to estimate the skill of each expert reduces the noise in the estimations.

However the impact of skill ratio on the standard deviation of the distribution changes from a two-expert combination to a multi-expert combination. In the case of

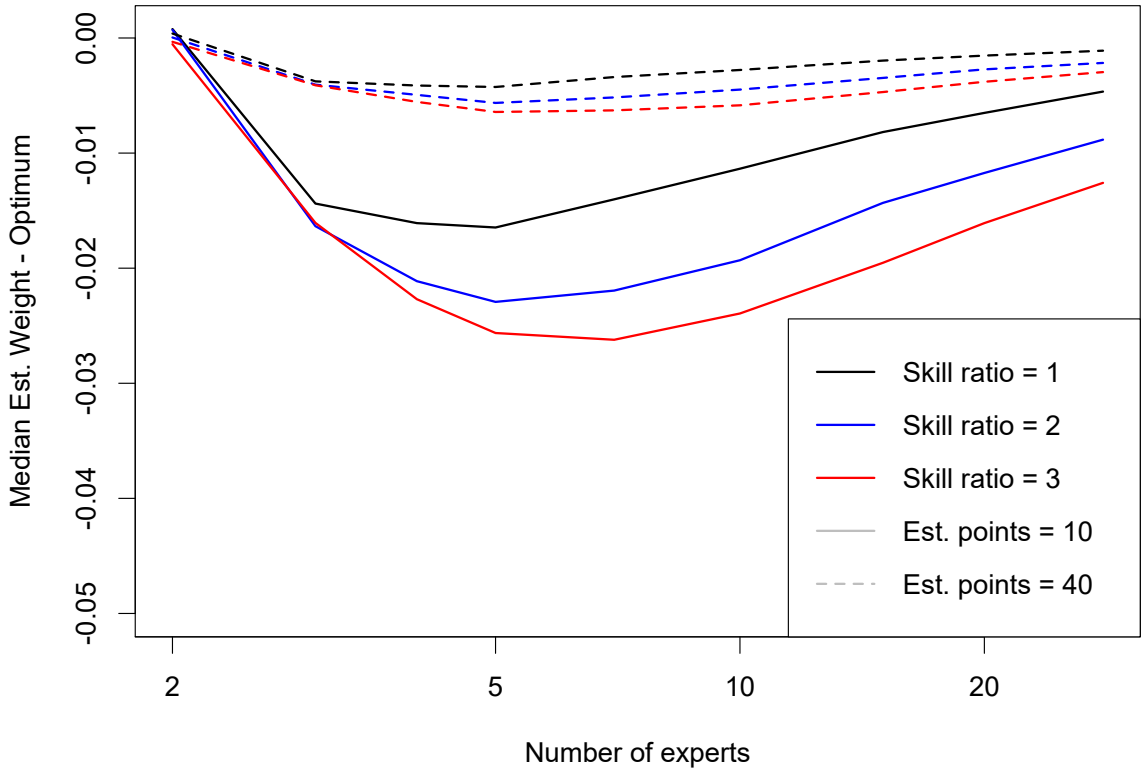


Fig. 8.: Impact of number of experts on the estimated weight bias (Median estimated weight minus variance optimal value) for various skill ratios and estimation sample sizes.

the two-expert combination, an increased skill ratio between the two experts creates a tighter distribution and lower standard deviation for the estimated weights. However, in multi-expert combinations with many experts (relative to the skill ratio of the most accurate expert), the skill ratio will increase the standard deviation of the weight as it acts as a scale factor in Equation 4.1. This is an interesting result as a higher skilled expert in a multi-expert system will increase the standard deviation of the estimated weight which in turn via Claeskens analysis (eq 2.9) can actually increase

the variance of the forecast combination.

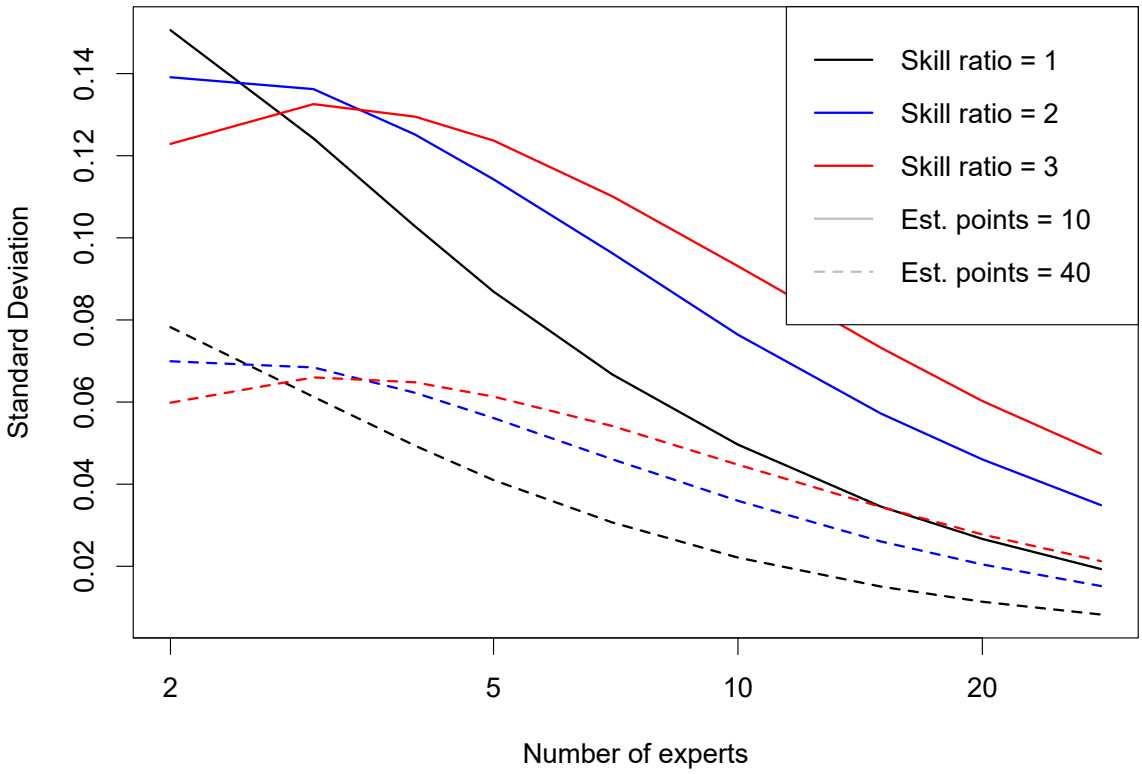


Fig. 9.: Impact of number of experts in the combination on the standard deviation of an estimated weight for various skill ratios and estimation sample sizes.

The complex interaction of the skill ratios and number of experts emphasizes the need to consider the shape of the weight distribution when deciding how to weight experts. The critical skill ratio approach described in Chapter 3 should be able to compensate for this and show improved combination performance in the range of 3-5+ experts. This interaction also suggests that findings from past studies of two-expert combinations may not be fully extensible to mulit-expert combinations. It is interesting to note that at 5 experts the number of experts is generally larger than the

max skill ratio one might expect to see in a combination of competent experts. Then at 5 experts, both the bias and standard deviation curves of an estimated weight begin to behave with the wisdom of the crowd effect where increasing experts dilutes these impacts. The dilution effect will only increase with expert combinations much above 5, while combinations with fewer than 5 experts will be more heavily influenced by the bias effects. This helps to explain the empirical observations that combinations of 5 experts tend to be more accurate than a simple average.

4.2.2 Impact of number of experts on critical skill ratios

From the previous section we have seen that increasing the number of experts in a combination can have significant impacts on the estimated weight distribution. We can use the same approach for estimating critical skill levels as described in Chapter 3 to determine the impact of multiple experts on forecast combinations. The definition of a critical skill ratio can be revised to use the multi-expert skill ratio as follows:

Definition 4.2.2. *Multi-expert critical skill ratio is the multi-expert skill ratio required such that the confidence level of using an estimated weight equals the decision maker's desired level of confidence.*

The estimated critical skill levels at a 50% level of confidence for multiple experts are displayed in Figure 10 based on the estimation process outlined in Chapter 3. The skill ratio needed for a given sample size decreases as more experts are added to the combination. As more experts are added, the sample size used to estimate the skill of the crowd in the denominator of the multi-expert skill ratio increases making the estimate of the average skill of the crowd more precise and making the overall estimate of the multi-expert skill level more precise. However, there are diminishing returns to the estimated skill ratio as even with many experts, the variability in the estimate

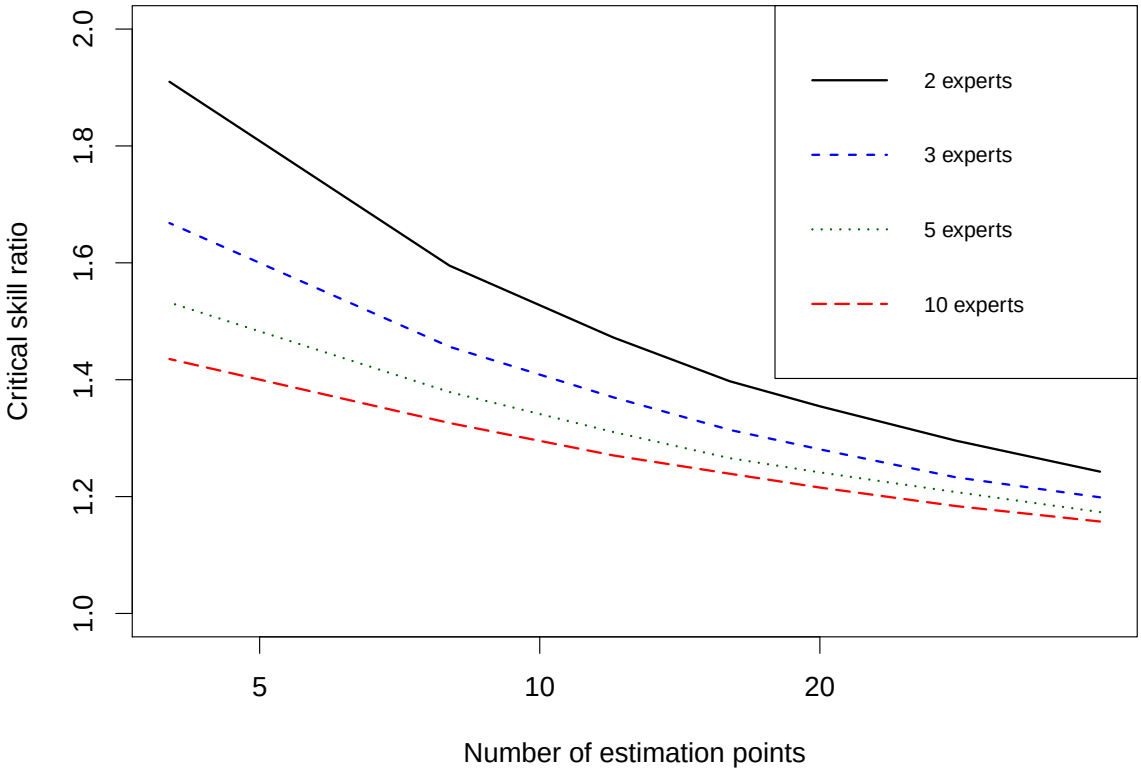


Fig. 10.: Multi-expert critical skill level evaluated at a 50% confidence for various expert combination sizes.

of expert 1 skill's remains constant for any number of experts.

In summary, it is clear that the distribution for estimated weights in a forecast combination is quite complex. When there are only two experts in the combination, the standard deviation of the weight distribution increases with skill ratio (Figure 9) and there is no bias in the distribution (Figure 8). When the number of experts is between 3 and $(2 + skillratio)$ the bias decreases as the number of experts increases while the standard deviation is nearly flat. When the number of experts is greater than $(2 + skillratio)$, we see the crowd effect kicking in and reducing both the bias

and the standard deviation. This suggests that there are three different regimes for forecast combination behavior to be considered. Much of the past research has focused on two-expert systems where the Equations are more tractable, but may not extend well to the other two regimes. The consideration of an outstanding (good or bad) expert versus a crowd of similar experts sheds light on the interaction of skill ratio and number of experts in a multi-expert system.

Although estimating weights and making decisions on approaches for combining forecasts may be complex, the use of the proposed critical skill ratio incorporates all of the complexity of the distribution in specific threshold values. These threshold values are a function of the number of experts and the number of sample points. It can be calculated once, in advance, and made available to decision makers to help guide their decision making. Based on the three regimes outlined, we would expect critical skill ratios to be most effective in the first two regimes where the interaction of expert skill and number of experts is most pronounced. As the number of experts increases when compared to the skill ratio of the most accurate expert, the dilution effect will tend to overwhelm any nuances in skill ratio interactions.

4.3 Improvements to forecast combination with multiple forecasts

We can use the critical skill ratio concept to better manage and understand when weight estimation error is likely to be significant and when it is not significant. This insight will help a decision maker address the two key questions raised in the literature review (Section 4.1):

- **Differential weights** When will skill based weight have a lower forecast combination error than a simple average?

- **Differential inclusion** Which experts should we include in a forecast combination?

We will use a simulation to test three new forecast combination heuristics (items 1-3 below) that leverage the critical skill ratio approach to answer the above questions. For comparison purposes we will compare these new proposed approaches against current approaches from the literature (items 4-7 below):

Approaches to be tested in the simulation:

1. **Best50 / Best90** Choose variance based weights (eq. 2.2) whenever one or more expert's estimated skill ratio is outside (high or low) of the 50% or 90% confidence critical skill ratio threshold for the given sample size and number of experts.

2. **Select90** Use variance based weights(eq. 2.2) on experts whose skill ratios are outside of the 90% confidence threshold (high or low) and average the remaining weight across experts within the thresholds.

3. **Drop90Avg / Drop90B50** Drop any expert whose estimated skill ratio is below the lower 90% confidence critical skill ratio threshold for the given sample size and number of experts. Then for Drop90Avg just average the forecasts of the remaining experts. In the case of Drop90B50, use skill ratio to make the choice for when variance based weights (eq 2.2) should be used on the remaining expert forecasts. If one or more experts are outside of the 50% confidence threshold, then use variance based weights. Use a simple average otherwise.

4. **Simple average (SA)** Arbitrarily give every expert an equal weight (eq 2.5).

5. **Variance optimal weights (Var)** Use variance based weights per Equation 2.2.

6. **AIC weights** Use the historical estimation sample to calculate the Akaike Information Criteria (AIC) for a simple average and for variance optimal weights (eq 2.2). Use the lowest value to decide whether to use a simple average or variance based weights. Note that in the original article, Schmittlein et al. (1990) also considered the selection of covariance based weights. For comparison purposes, we are limiting the choice to the same set of choices that Best50 and Best90 are making. In this simulation, the experts will be modeled with no covariances.

7. **Top5** When there are more than 5 experts, choose only the top five based on historic mean absolute error performance and take a simple average.

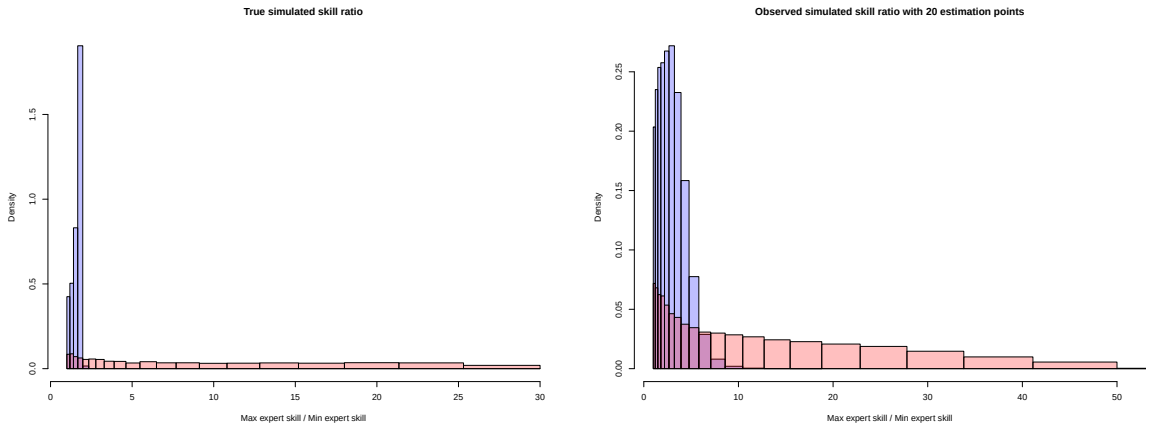
The Best50 or Best 90 approach manages the impact of estimation error by only estimating weights when there is an expectation that estimated weights will perform better than a simple average. The decision maker can choose the level of risk of making a wrong decision by varying the desired confidence level. At 90% confidence the heuristic will be more biased towards choosing a simple average, while a 50% confidence level will more frequently capture improvement opportunities when skill based weights are preferred. However a 50% confidence level may also allow more wrong decisions which can increase the errors. Select90 is a more nuanced approach that may further reduce estimation error by only estimating weights for those few experts that are significant outliers (good or bad) and not taking on additional estimation error for experts with relatively undifferentiated skill. The Drop90 then average the remainder approach addresses the dilution effect of a crowd by eliminating those experts who are significantly less accurate than the crowd. In the case of truly independent experts (ie no covariance), one would expect that keeping every expert would be beneficial as each expert brings some new information to the combination. But with more ex-

perts, there is a greater opportunity that one of the expert's skill is poorly estimated. The work on only choosing the Top5 experts recognizes the diminishing returns of adding experts and suggests that 5 is the optimal number of experts. Using critical skill ratio to drop experts that are significantly less skillful may be a better approach than taking the Top5. Both of these methods for dropping experts may have stronger results when we consider the impacts of expert correlations in Chapter 5. Finally, even after dropping experts, there may be a wide enough dispersion in the skill of the remaining experts to justify using skill based weights on the remaining forecasts. The Drop90B50 uses 50% critical skill threshold evaluated for the number of remaining experts to make this choice. Drop90B50 may provide an additional performance benefit when there are wide-ranging skills across the experts.

4.3.1 Simulation methodology

A decision maker may be aggregating forecasts during a period of relative stability (economic, weather, political) where data and assumptions to make a forecast are likely to be more similar and the relative range of expert forecasts and errors are similar. This is a low dispersion environment as describe by Mannes et al. (2014). Alternatively, the aggregation may be during a period of relatively more instability in conditions, data or assumptions being used resulting in a larger range in the forecasts and errors of each forecaster. The decision maker may not know, a priori, the stability in the underlying data and assumptions each of his experts is using, nor know the likely future stability in the quantity being forecast. Based on this, we will assume that the decision maker does not have a strong prior belief about the relative dispersion of the forecasts or the skill of the experts in his combination. Therefore, we will simulate an average of both environments that is a 50% sample from a distribution that is low in forecast dispersion and expert skill dispersion and a 50% sample from

a distribution that is high in expert forecast and skill dispersion to see which aggregation methods are broadly effective. Figure 11 illustrates the relative range of skill in the combinations low dispersion and high dispersion environments as measured by the skill of the most accurate expert divided by the skill of the worst expert.



(a) True skill distribution (b) Observed distribution

Fig. 11.: Range of experts skill used in simulation samples as measured by the skill of the most accurate expert / skill of the worst expert. The observed distribution was based on 20 point estimation samples.

We used a simulation approach similar to that of Mannes et al. (2014) except that in this chapter we assumed no correlations between the experts. We looked at combinations of 2,3,5,7,10,15,20,28 experts to examine the impact of increasing numbers of experts. For each of the k experts in a combination, the expert's mean absolute error MAE_i was sampled from a uniform(83,117) distribution in the low dispersion case and from a uniform(31,169) in the high dispersion case. Then for each expert, 70 points were sampled from a Normal(0,1) distribution multiplied by $\sqrt{\frac{\pi}{2}}MAE_i$. In this approach, the resulting low dispersion expert true error variance will range 10,821 to 21,503 or a 2:1 ratio. For the high dispersion distribution, the

expert true error variance will range from 1,510 to 44,863 or a 30:1 ratio.

For each combination of k experts sampled, we looked at estimation sample sizes of $n \in (4, 8, 12, 16, 20)$ reflecting 1 year to 5 years of quarterly data. On a rolling basis the first n of the 70 sample points were used to estimate weights for each of the decision rules under exploration and applied to the simulated errors at the $n + 1$ sample point. This was repeated 50 times on a rolling basis by incrementing the beginning of the estimation sample. Then for each sample the mean absolute error was calculated across the 50 iterations for each rule. This overall process was repeated 1000 times for each number of experts k using the low dispersion skill distribution and then another 1000 times using the high dispersion skill distribution.

To compare each aggregation method for a particular combination of experts and estimation points, we normalized the MAE for each decision rule by calculating the percentage improvement for each rule over the average expert's mean absolute error at that sample point. In this case, if a particular sample had a very large average error such that some decision rules show large improvements, this one sample won't over-shadow other samples where the average expert error may be smaller but the percentage gain from using a particular rule is high. The mean improvement is then calculated for each rule across the 2000 sample points for the number of experts and each estimation sample size. The decision rule with the largest gain in mean absolute percent improvement (MAPI) is considered the optimal choice. Using mean squared errors generated similar results.

4.3.2 Choosing the best approach - simple average or skill based weights

The performance of a simple average (SA) and various skilled based estimated weight approaches is provided in Table 4. A simple average performs better than both variance based weight (Var) and AIC weights (AIC) when there are only 4 points;

however as more points are available, these methods are able to perform better than SA. With few estimation points skill estimation errors will be higher favoring SA. The performance differential is not substantially impacted by the number of experts but as more experts are considered the overall performance improvement increases as more independent experts provide more information.

The use of a critical skill ratio threshold (CSKr) to decide if a variance based weight or a simple average should be used (methods Best50 & Best90) performs better than all three of the benchmark methods (SA, Var, AIC) with only $4-8$ points. As the number of estimation points increases the uncertainty around what is the most accurate method and what is the correct skill is reduced which diminishes the advantages of the critical skill ration approach relative to other methods. At 20 estimation points the critical skill ratio approaches are either equal or slightly better then the three benchmark methods. Setting the critical skill ratio threshold at a 90% confidence level (Best90) only outperforms a 50% level (Best50) when there are many more experts. As the number of experts increases,the random probability that one of the expert's skill has been errantly estimated above the threshold will increase. This mis-estimation would bias Best50 to select a variance based weight in error more frequently. Setting a threshold of 90% (Best90) when there are more experts compensates for this effect. Overall Best90 is the most accurate performer, or nearly so, when there are 10 or more experts. It is interesting to note how well a decision rule (B50/B90) that relies on estimates performs with only 4 estimation points and few experts. With five or fewer experts we are working in the regime where skill level is on par or dominates the number of experts in the weight estimation process (eq 4.1). This is the area we expected our rule to be most effective. In general the rule performs well with 5 or fewer experts, as the size of the crowd reduces relative performance advantage.

	Mean percent improvement over Avg. expert error								
	4 est. points			8 est. points			20 est. points		
Num. Experts	3	10	28	3	10	28	3	10	28
SA	40.7	67.1	80.4	40.7	67.1	80.4	40.7	67.1	80.4
Var	39.2	63.3	75.6	43.1	68.4	80.9	45.5	71	82.8
AIC	40.6	66.8	80.4	43.1	67.9	80.4	44.9	69.6	80.9
Best50	**42.6**	**68.4**	80.9	**44.3**	**69.9**	82.1	45.5	71	82.8
Best90	41.8	68	**81.2**	43.3	69.8	**82.3**	**45.6**	**71.1**	82.8
Select90	41.8	67.4	80.4	43.2	68.4	81	**45.6**	**71.1**	**82.9**

Table 4.: Mean average percentage improvement (⇑ better) over an average expert for various forecast combination methods based on simulation results for various combinations of experts and estimation points.

4.3.3 Selectively choosing which experts to use skill based weights

A more nuanced approach to reduce estimation error is to selectively estimate the weights of only those experts that have clearly differentiated performance (good or bad) and use a simple average for the remaining weights. Table 4 hows that a selective approach at a 90% level of confidence (Select90) is, on average, always as good or better than the three benchmark approaches. However, Select90 is only marginally better than Best90 or Best50 when there are many experts and many estimation points to accurately assess skill. This approach does not appear effective in smaller combinations but may be worth revisiting in combinations with many more than 28 experts as it should improve with more experts.

4.3.4 Choosing which experts to include

In addition to differentially weighting experts, strategies to differentially include experts have been proposed to offset the dilution effect of many experts in a combination. Using the critical skill ratio at a 90% level of confidence to decide which experts to include and which to drop is proposed as a possible improvement over arbitrarily retaining the Top5 experts recommend by Mannes et al. (2014). Both Drop90Avg and Drop90B50 consistently perform better than Top5 in all scenarios (Table 5). This result is not entirely surprising as we have purposely simulated all experts as independent. The Drop90Avg and Drop90B50 generally retain more experts than Top5 and are therefore expected to do better with independence. For the same reason SA consistently does better than Top5 given the independence assumption in the simulation. We will revisit this result in Section 5.4 when we run a simulation with expert error correlations.

It is interesting to note that Drop90Avg performed as well as or better than SA despite having fewer experts. Again with more independent experts in its combination, one would expect SA to perform better than Drop90Avg. In this case, Drop90Avg finds a more optimal number of better experts such that the lower dilution effect on the higher skilled experts is sufficient to stay even with SA. Drop90B50 consistently performs better than SA and Drop90Avg. In this case, the decision to use variance based weights when over the critical skill ratio enhances performance, similar to the results of Best50 versus SA. Even so, Best50 is consistently a better approach than any of the inclusion methods in a no expert correlation environment. Both of these results are promising and suggest that a critical skill ratio approach may be a better rule for determining expert inclusion. We will revisit this with expert error correlation in Section 5.4.

	Mean percent improvement over Avg. expert error								
	4 est. points			8 est. points			20 est. points		
Num. Experts	3	10	28	3	10	28	3	10	28
SA	40.7	67.1	80.4	40.7	67.1	80.4	40.7	67.1	80.4
Best50	**42.6**	**68.4**	**80.9**	**44.3**	**69.9**	**82.1**	**45.5**	**71.0**	**82.8**
Top5	NA	59.7	64.2	NA	60.8	66.7	NA	61.9	68.2
Drop90Avg	41.0	67.1	80.4	41.7	67.4	80.4	40.9	67.4	80.4
Drop90B50	42.4	68.2	80.8	43.9	69.7	82.0	41.2	68.7	81.3

Table 5.: Mean average percentage improvement over an average expert (⇑ better) for various forecast combination methods that compares SA and Best50 to methods which exclude experts. Based on simulation results for various combinations of experts and estimation points.

4.4 Discussion

The simulation was designed to address to questions:

- **Differential weights** When do estimated weights yield a lower error than a simple average?
- **Differential inclusion** Which experts should we include in a forecast combination?

The performance of the Best90/ Best50 decision rules demonstrates that it is possible to improve upon a simple average. The rules guide a decision maker when there is sufficient differentiated skill level to justify using differentiated weights. In the worse case that the experts are not highly differentiated, both rules default to using a simple average. The advantage of these rules decreases as the number of

experts increases and dilution effects become dominate. These rules are based on the concept and estimates of a multi-expert critical skill ratio that is based on the harmonic average of the crowd versus one expert. As expected from the analysis, we did see the strongest performance of the decision rules in the 3-expert range where the interaction between skill level and number of experts is greatest.

The performance of Drop90Avg and Drop90B50 demonstrate that choosing experts based on the 90% critical skill level was better than including just the Top5 experts. Although not entirely surprising in a no covariance environment, these results are encouraging as the selection of a few experts was able to perform as well as a crowd of all experts.

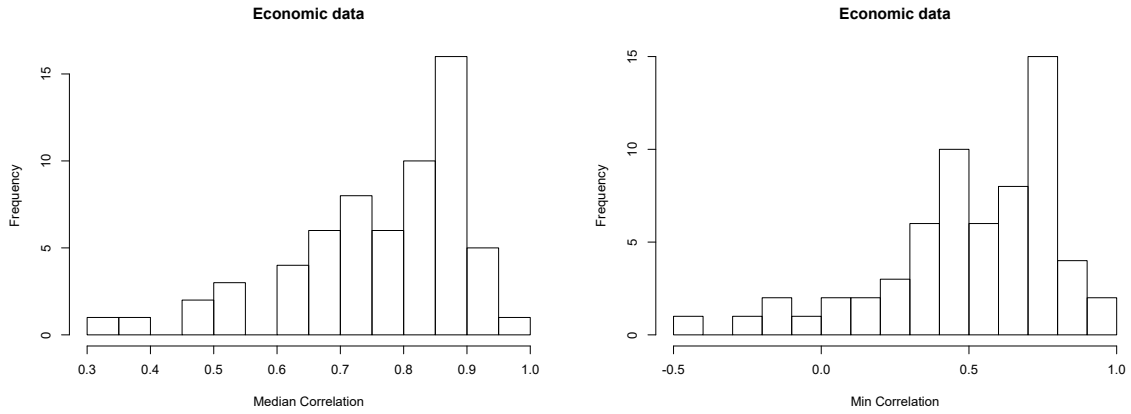
In Chapter 5 we will analyze the impacts of correlations between experts and revise the decision rules for expert correlations. Once we have decision rules that account for expert correlation, we can compare them to actual expert forecasts where expert correlations are frequently found (Chapter 6).

CHAPTER 5

COMBINATION OF MULTIPLE FORECASTS WITH CORRELATED EXPERTS

In the previous chapters we have assumed no correlation between the expert forecasts. As experts often draw on the same or similar historical data and may use similar methods and models, it is highly likely that the forecasts and forecasts errors are correlated to some degree. A review of 78 economic forecasts from G7 and emerging market countries (see Section 6.1.1 for details) shows that the median expert pair correlation is frequently greater than 0.7 and the minimum expert pair correlation is above 0.3 at least 85% of the time (see Figure 12). These estimates of expert correlations are similar to those found by Clemen (1986) for different data series and time windows.

In this chapter we will use the multi-expert skill ratio approach from Chapter 4 to analyze multi-expert forecast combinations with expert correlations. This analysis will help explain why covariance weights are sensitive to estimation errors and therefore, in practice, not effective for forecast combinations. Insights from this analysis will be used to develop a new heuristic for estimating covariance based weights that performs better than a simple average in both simulation studies and in a large body of real forecasts. In addition, we will extend the development of the critical skill ratio threshold heuristic to forecast combinations with expert correlations.



(a) Median expert pair correlation (b) Minimum expert pair correlation

Fig. 12.: Observed expert pair forecast error correlations based on 78 quarterly forecasts of 7 economic indicators from 15 countries as collected and published by a private economics consulting firm (see Section 6.1.1 for details).

5.1 Literature review - Combination of multiple experts with correlated experts

The sensitivity of weights due to expert correlation has been examined by a number of researchers. Kang (1986) found considerable variability in the period-to-period estimation of covariance optimized weights for 4 expert forecasts of quarterly US nominal GNP growth from 1970 to 1982 . In this study, the previous 20 forecast periods were used to estimate the covariance optimal weight per Equation 2.1. They found that the weight of any one expert could change by as much as 1.0 out of a total weight of 1.0 across all experts. The standard deviation of the period-to-period change was above 0.20 for each of the 4 experts. Gunter (1992) performed a similar study on 5 time series forecasts for quarterly US GNP growth from 1952 to 1987 and found that the 90th percent quantile period-to-period change in an ordinary least squares estimated weight with 21 data points was 1.11 In a simulation study,

Smith and Wallis (2009) also found a wide dispersion in estimated weights. The observed size of the estimated weight change in one period of data calls into question the practicality of these theoretical optimal approaches. If the data is indeed that variable, then the 20 periods of history used to estimate the covariance matrix is not likely to be representative of the current period.

Winkler and Clemen (1992) used the Pearson Type VII distribution to estimate the standard deviation of an estimated weight with expert correlations. They found that in a two-expert combination, weights are most sensitive to changes when the skill ratio is near 1.0 or when the expert correlation is near 1.0. They extended their approach to multiple expert combinations by estimating the combined variance of two experts and then stepwise, adding a new expert to the previous average until all experts have been combined. They observed that as the number of experts increases, the effective expert skill ratios and correlations estimated by their sequential method came closer to the region of instabilities for a two-expert combination. Winkler and Clemen (1992) estimated the standard deviation of an estimated weight to be above 0.2 when expert correlation is above 0.8 and skill ratios above 1.2 in a two-forecast combination with 30 estimation points. This estimation error alone is sufficient to explain the weight variations observed by Kang.

Blanc and Setzer (2016) used Winkler and Clemen's work on standard deviation of estimated weights to explore the critical sample size in a two-expert system where the loss in accuracy due to estimation error of covariance estimated weights is equal to the gain in accuracy from using estimated weights over fixed weights of a simple average. They found that a smaller sample was required as the relative skill of the two experts diverged (larger skill ratio) and that a smaller sample was required as the amount of correlation between the two experts increased. The thresholds and sample sizes developed theoretically by Blanc closely matched those empirically determined

by Schmittlein et al. (1990). Blanc also looked at how robust the required sample sizes were to structural changes in the experts' skill, and correlations. They found that the robustness of covariance based weights to structural changes decreased with increasing expert correlations, consistent with the general observation that increasing correlations creates instability in weights.

The estimation of the expert error covariance matrix has been identified as a key contributor to the poor performance of covariance optimal weights (Blanc and Setzer, 2016; Bunn, 1985; Chan and Pauwels, 2018; Clemen, 1986; de Menezes and Bunn, 1998; Smith and Wallis, 2009). Clemen (1986) investigated ways to improve upon the estimation of the covariance matrix with economic forecast data. They found that estimating parameters with decreasingly weighted data points was not an effective approach. They proposed a Bayesian estimation approach that used an inter-class matrix with equal expert variances and covariances as the prior distribution. This prior information stabilized the estimate of the covariance matrix and approached the performance of a simple average. This improvement is achieved by basically shrinking the estimated weights towards fixed weights in a simple average.

Researchers have recognized the presence of high expert correlations and have explored weight estimation methods that are more robust to correlations. Guerard and Clemen (1989) explored the use of latent root regression to remove non-informative data collinearities in the estimation process. They found that out-of-sample performance for latent root regression was not significantly different from ordinary least squares regression (OLS), nor did it match the performance of a simple average. Genre et al. (2013) explored using multiple methods including principle component analysis (PCA) on clusters of expert forecasts of GDP growth, inflation rate, and unemployment rate from the European Central Bank Survey of Professional Forecasters. PCA performed poorly on GDP and unemployment forecasts as compared to a simple

average. It did beat a simple average on inflation forecasts, however in this case, it did not beat a naive forecast of inflation due to the quality of the forecasts. Ordinary least squares were also used to estimate weights for clusters of similar experts. The case with three clusters, no constant term, and weights constrained to sum to one, outperformed a simple average on a one-year time frame but not on a two-year time frame. Other combinations of clusters and parameter constraints did not out perform a simple average.

Rather than changing the estimation method, adding non-negativity constraints on expert weights have been found to improve the stability and performance of forecast combinations. Variance optimal weights are naturally constrained to be positive by construction and as previously referenced, are generally found to outperform covariance optimal weights. Gunter (1992) found that the 90th percentile period-to-period change in weights estimated by non-negatively constrained ordinary least squares was only 0.25 as compared to 1.11 for unconstrained ordinary least squares. In a more extensive study with 40 time series, Aksu and Gunter (1992) found that non-negatively restricted least squares combination models most frequently outperformed other ordinary least squares methods and matched a simple average. Genre et al. (2013) found that constrained OLS performed better than unconstrained but was not better than a simple average in two of their series. Conflitti et al. (2015) used a LASSO regression approach to constrain the weights to add to 1.0 and to all be positive. Combination of ECB SPF forecasts of GDP growth and inflation rates using these constraints performed slightly better than a simple average but not statistically different. An advantage of the LASSO approach is that the constraint will force the selection of some but not all of the experts. In this case, between 1 to 6 experts were retained which was quite small when compared to the available number of experts.

Another impact that expert correlation can have is to reduce the effective in-

formation available to the decision maker Clemen (1987). If each expert has some unique information but also has some information in common with other experts, the correlations that the common information induces makes it harder for the decision maker to assess the unique information of each expert. In this case, higher correlations, or more experts with correlations, can have a negative impact on the estimated weights in a forecast combination. In a more narrow case, if all of the experts have the same skill and correlations between experts are the same, then the effective number of independent experts in the combination is limited to ρ^{-1} the inverse of the common correlation level (Clemen and Winkler, 1985). In the case of a common expert correlation of 0.5, which is low based on the observations in Figure 12, the effective number of experts in the combination is only 2 regardless of how many experts are added. These observations provide a theoretical basis for using a small subset of experts in a combination when there is correlation.

In summary, although covariance weights minimize the in-sample combination variance, the variability in the estimates results in poor out-of-sample performance. Strategies that constrain the weights and remove experts have had the most success offsetting the estimation issue but have not been able to confidently beat a simple average. In the next section we will extend the analysis of the impact of expert correlations from a two-expert combination to a multi-expert combination. Insights from this work will be used to develop an improved heuristic for estimating covariance based weights that overcomes the estimation issue. This heuristic will be tested in a simulation study (Section 5.4) and with real data (Chapter 6).

5.2 Impacts of expert correlation on estimated weights

In Chapter 3 we introduced the skill ratio of the most accurate expert versus the crowd to tease out the impacts and interactions of expert skill and size of crowd on

the optimal variance based weights (Equations 4.1 and 4.2) for aggregating experts. For the analysis of covariance impact, we will assume that all of the experts share the same, positive degree of correlation. This assumption is a much closer reflection of reality than assuming no covariance and will provide insights to how covariance impacts the overall forecast combination with multiple experts. We will further assume that each member of the crowd (not including expert 1) have the same variance. This is assumption is a bit more restrictive than using the harmonic average of the crowd but it facilitates the analysis. We will test our findings by simulating experts who do not have the same level of correlation or the same level of variance to see whether the findings generally still hold when these assumptions are violated.

Then, in this simplified case, the expert forecast error covariance matrix $\mathbf{\Sigma}$ can be formed as follows from three variables: $\sigma_{crowd}, Skr_1, \rho$ where ρ is the common correlation between expert pairs. This approach is similar to the one used by Clemen and Winkler (1985) except that it determines how skill ratio and correlation interact

in the determination of an optimal weight.

$$\boldsymbol{\Sigma} = \mathbf{SAS}$$

$$\text{Where: } \mathbf{A} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} \frac{\sigma_{crowd}}{\sqrt{Skr_1}} & 0 & \dots & 0 & \\ 0 & \sigma_{crowd} & \dots & 0 & \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & \dots & 0 & \sigma_{crowd} & \end{bmatrix}$$

$$\text{then: } \boldsymbol{\Sigma}^{-1} = \mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S}^{-1}$$

The inversion of the $\mathbf{S}$ and $\mathbf{A}$ both have known solutions which can be used to evaluate $\boldsymbol{\Sigma}^{-1}$. We can then use the optimal covariance based weights approach (Eq. 2.1) to determine the optimal weights in the presence of a common expert error correlation (see Appendix D):

$$w_1 = \frac{Skr_1 + \rho[(k-2)Skr_1 - (k-1)\sqrt{Skr}_1]}{Skr_1 + (k-1) + \rho[(k-2)Skr_1 - 2(k-1)\sqrt{Skr_1}]} \tag{5.1}$$

$$w_{crowd} = \frac{1 - \rho\sqrt{Skr_1}}{Skr_1 + (k-1) + \rho[(k-2)Skr_1 - 2(k-1)\sqrt{Skr_1}]} \tag{5.2}$$

When $\rho = 0$, Equations 5.1 and 5.2 simplify to Equations 4.1 and 4.2 previously developed for the no-covariance scenario. When $\rho = 1$, the $\mathbf{A}$ matrix becomes singular and is not invertible. When ρ is close to 1 the determinant of matrix $\mathbf{A}$ becomes very small making the weights very sensitive to estimation errors. When $Skr_1 = 1$, Equations 5.1 and 5.2 simplify to equal weights $\frac{1}{k}$ for all ρ.

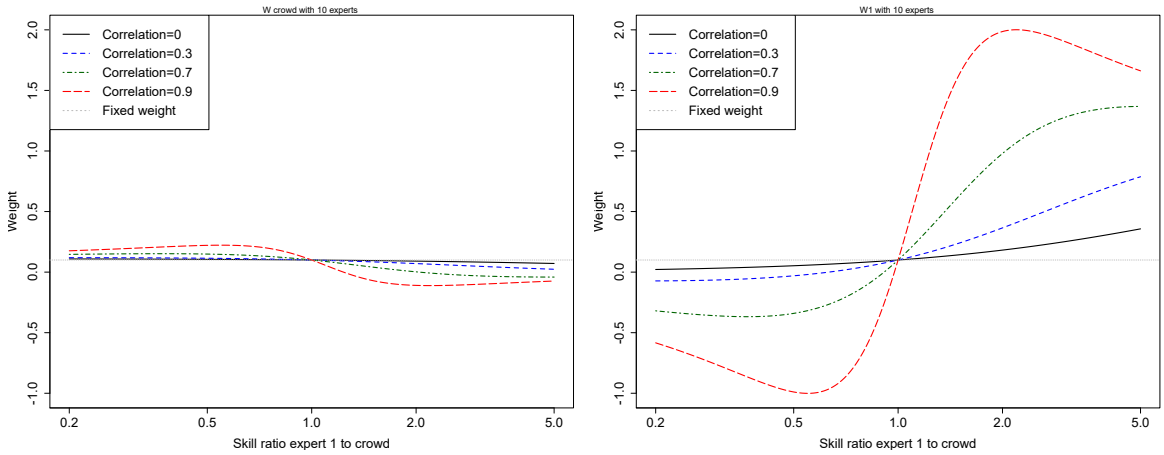
When $k = 2$, Equations 5.1 and 5.2 simplify to the covariance based weights for a two-expert system often cited in the literature where the expert correlation is only multiplied by the square root of the expert skill ratio (Bates and Granger, 1969; Winkler and Clemen, 1992):

$$w_1 = \frac{Skr_1 - \rho\sqrt{Skr}_1}{Skr_1 + 1 - 2\rho\sqrt{Skr_1}} = \frac{\sigma_2^2 - \rho\sqrt{\sigma_1\sigma_2}}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$
$$w_2 = \frac{1 - \rho\sqrt{Skr_1}}{Skr_1 + 1 - 2\rho\sqrt{Skr_1}} = \frac{\sigma_1^2 - \rho\sqrt{\sigma_1\sigma_2}}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

However, when there are more than two experts, the expert correlation in the numerator of weight of expert 1 is multiplied by an additional $(k-2)Skr_1$ term which will increase the impact of expert correlation on the weight of expert 1 as the number of experts increase and as their relative skill increases. In contrast, the numerator for the average weight of the crowd is only impacted by the square root of the skill ratio with no corresponding impact from the size of the crowd. The impact of expert correlation on covariance optimal weights using Equations 5.1 and 5.2 is illustrated in Figure 13A for a member of the crowd and 14B for an expert with a skill ratio different from the crowd. In both cases, the weights converge to a focal point of equal weights when the skill ratio is 1. However, the weight of the outlier, expert 1, is significantly more sensitive to skill ratio than the crowd is. This sensitivity (slope of the $w_1 \propto log(Skr_1)$ line), increases rapidly with increasing levels of expert correlations. Another asymmetry is in the weight assigned to an outlier in Figure 13b. With a correlation level of $\rho = 0.7$, an expert who is twice as skilled as the crowd will receive a weight of 0.98 while an expert who is one-half as skillful as the crowd will receive a weight of only -0.34. This asymmetry is not present in two-expert systems due to only 1 degree of freedom, but becomes significant in a multi-expert system. Basically, when experts tend to say the same thing, there is a much higher premium on the

more accurate expert and all of the others become less differentiated as a result.

The amount of weight expert 1 receives is highly impacted by the level of expert correlation. Even at low correlation levels $\rho = 0.3$, the weight of expert 1 is approximately twice the weight assigned by variance based weights $\rho = 0$. At very high levels of expert correlation, the relationship appears more like an on/off function. For example with a 20-expert combination and a 0.9 common correlation between experts, expert 1's weight becomes zero at a skill ratio of only 0.81, while at a skill ratio of 1.24, expert 1 will receive a weight of 1.0. We will explore the point at which experts receive zero or negative weights shortly.



(a) Weight of an average crowd member (b) Weight expert 1

Fig. 13.: Covariance optimal weights (Equations 5.1 and 5.2) for various levels of expert correlation in a 10 expert combination.

The differences in how the three methods assigning weights (simple average, variance based, covariance based) can be seen in Figure 14. The simple average assigns the same weight to all experts regardless of skill level, so that the relative weight is always 1.0. Variance based weights assign relative weight that is proportional to the skill relative skill level with a slope of 1.0 (Equations 4.1 an 4.2). Covariance

based weights assign weight to the better expert at a higher rate than variance based weights do. In this case, the slope of the relative weight line is not linear and increases with skill ratio. The non-linearity is most pronounced as the degree of expert correlation increases. This non-linearity coupled with estimation errors help explain the high variability observed in covariance weights (Gunter, 1992; Winkler and Clemen, 1992). It is interesting to note that for covariance weights, the weight is assigned asymmetrically. Higher skilled experts will be assigned relatively more weight, than that assigned to a correspondingly lower skilled expert.

The impact of increasing number of experts on covariance optimal weights in the presence of expert correlations is illustrated in Figure 15. The weight curves for expert 1 all converge at $w_1 = 1.0$ when the $Skr_1 = \frac{1}{\rho^2}$ and the average weight of the crowd is 0 per Equation 5.2. Prior to this inflection point, more experts decreases the weight of expert 1, consistent with the dilution effect previously discussed in Chapter 4. In this region, an increasing number of experts has a small positive impact on the slope of the $w_1 \propto log(Skr_1)$ line but not nearly to the degree of impact that increasing expert correlation does. It is interesting to see how linear the $w_1 \propto log(Skr_1)$ relationship is between weights of 0.0 and 1.0. The slope of the line appears to be a function of the number of experts and level of common expert correlation. In the region beyond the inflection point, expert 1's weight is greater than 1.0 and actually increases with increasing number of experts. With very high weights, the accuracy of the combination is mostly determined by the accuracy of the highly-skilled expert without much of a "crowd" benefit. If this one expert's skill has been estimated in error, then the accuracy of the overall combination is at risk. This suggests forecast combinations that assign very high weights may not be optimal in practice. We will now look more closely when expert 1's weight is very high and the corresponding weights of the crowd may be negative.

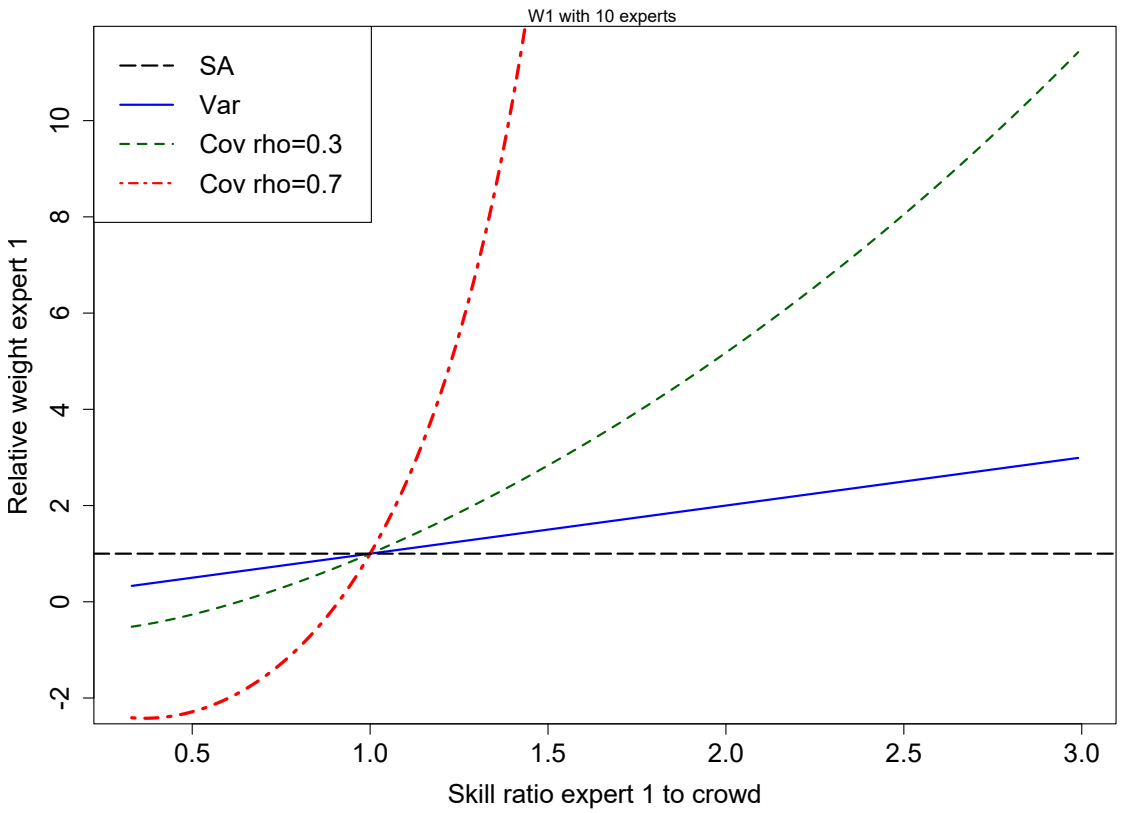


Fig. 14.: The degree of differential weighting each weight estimation method assigns to a higher skilled expert for various levels of skill ratio for a 10-expert combination. For covariance weights equation 5.1 and 5.2 are used with a common correlation assumption.

5.2.1 Impact of negative correlations

Empirical data suggests that experts are typically positively correlated due to shared models and data (Elliott, 2011; Winkler and Clemen, 1992), however for completeness and curiosity sake, one might ask, how do weights behave when all of the experts share a negative correlation with each other? Figure 16 shows the impacts of increasing negative correlations. In this case the weights are not nearly as

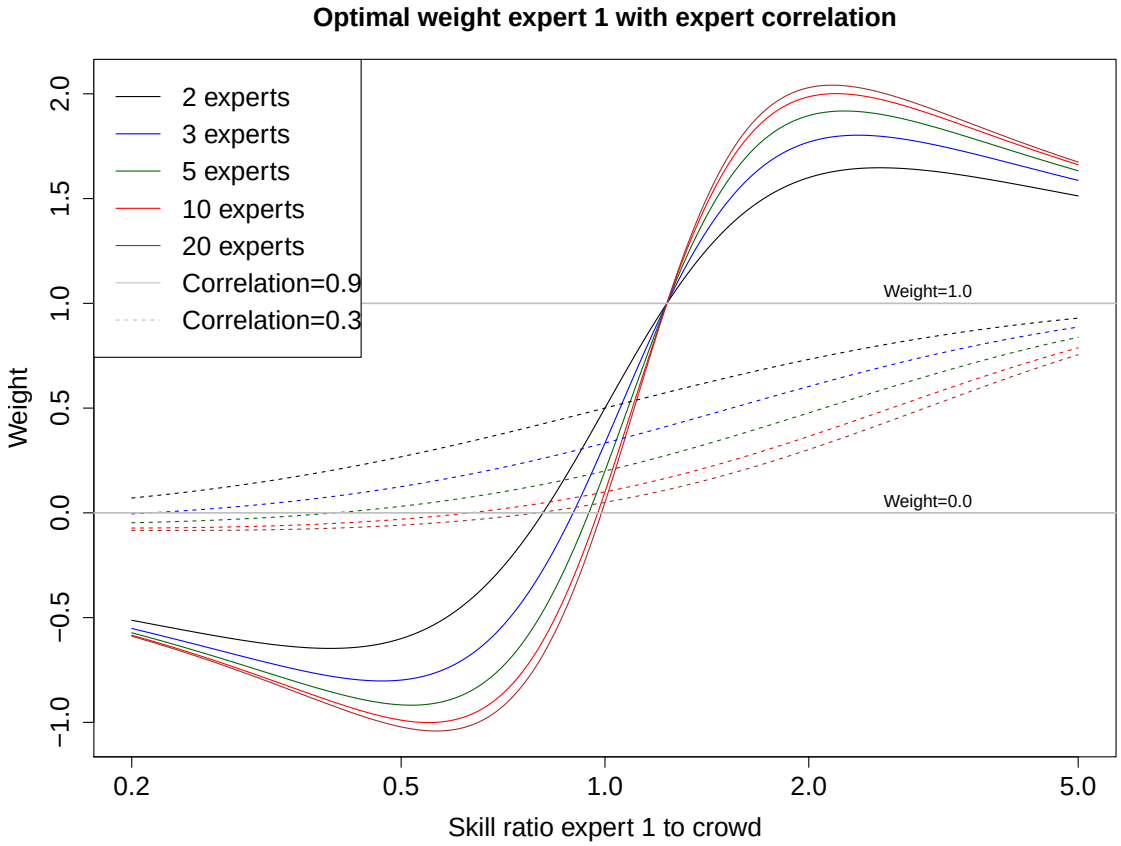
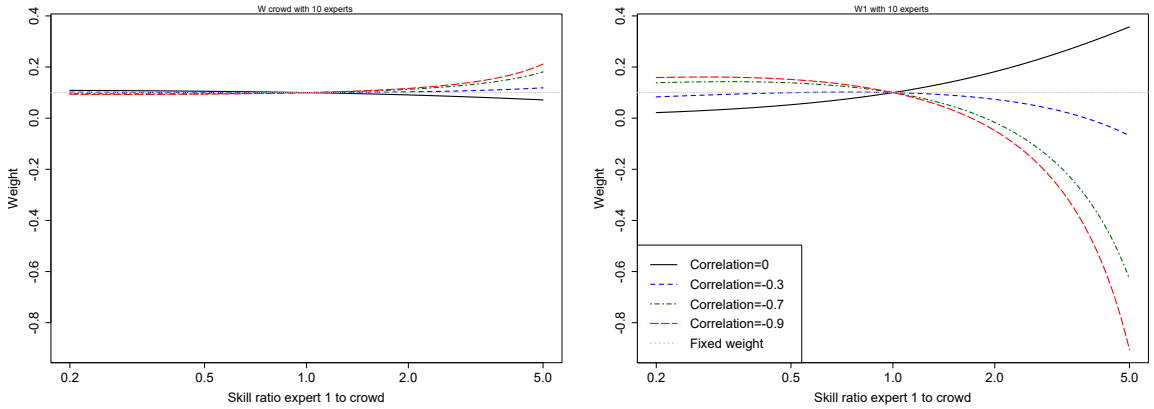


Fig. 15.: Covariance optimal weights (Equations 5.1 and 5.2) for various numbers of experts in the combination

extreme for a corresponding skill ratio. In the case of negative weights the forecasts will naturally offset themselves so an optimum combination can be obtained without high extremes in relative weights.



(a) Weight of an average crowd member (b) Weight expert 1

Fig. 16.: Covariance optimal weights (Equations 5.1 and 5.2) for various levels of negative expert correlation in a 10 expert combination.

5.2.2 Expert correlation and negative weights

With equal weights and with variance optimal weight, no expert receives a zero weight regardless of their skill. Based on the independence assumption each expert has some information to contribute. However, when we introduce correlation between experts this may no longer be true. If expert A directionally repeats what expert B says but only with proportionally more magnitude (gusto) then she/he is not necessarily adding any new information as there is only a scale difference between the two forecasts. In this situation, there will be a point of correlation where it makes more sense to just use the most accurate expert and ignore the one who consistently overshoots in the same direction as the better expert. Taken to an extreme, if we can rely on the consistency of the strong expert correlation, the degree of overshoot can be used to predict and then correct the size of error of the better expert. In this situation, the lower-skilled expert is given a negative weight. This scenario only works if there is a strong confidence in the high degree of expert correlation which is

rarely the case. At high levels of expert correlation, the determinant of the expert correlation matrix approaches zero and the estimated weights become very sensitive to errors. Hence, the poor performance of covariance weights and negative weights previously observed in the literature.

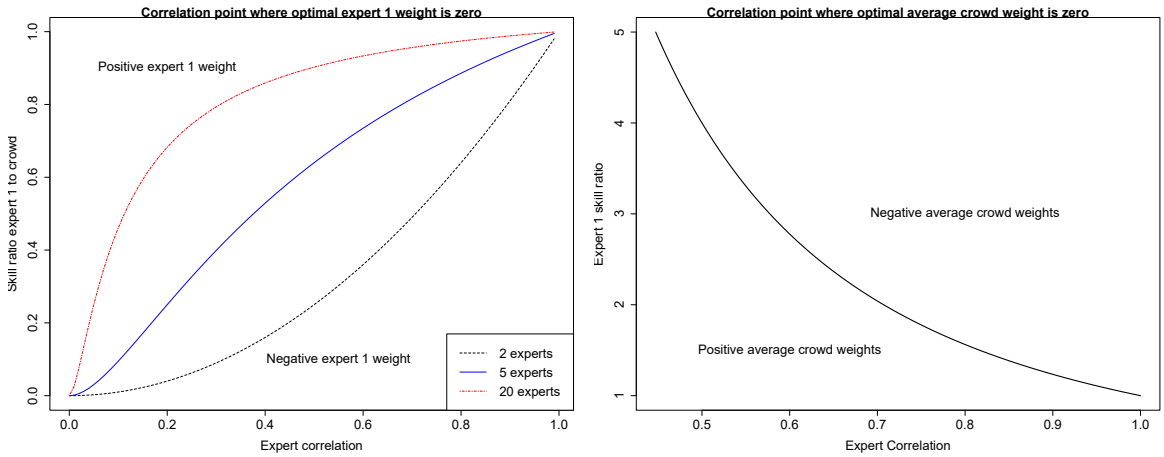
The proposed covariance optimal weights (Equations 5.1 and 5.2) both permit zero and negative weights. At high levels of expert correlation and when the skill of expert 1 is lower than the crowd, there comes a point where the expert is saying, directionally, the same thing as the crowd but with less accuracy and therefore is not adding to the overall accuracy. Conversely, when there is a high degree of correlation and the expert is better skilled than the crowd, the crowd will be saying directionally, the same thing as the expert but with less accuracy and here too is not adding to the overall accuracy. In both cases, the overall sum of weights is still constrained to one. So if the average weight of the crowd is 0.0 then the weight of expert 1 will be 1.0 and vice versa. Note that Equation 5.2 is the weight of crowd members who all have the same skill, and for this discussion it can be interpreted as the average weight of the crowd. Even when the optimal average weight of the crowd is zero some more skillful members of the crowd may have a positive weight offset by other less skillful members who have a negative weight.

Figures 18(a) and (b) illustrate the expert correlation and skill ratio relationships where the weight of expert 1 is zero and where the average weight of the crowd is zero. When there are 5 experts in the combination, that point at which expert 1's weight is zero is nearly a linear relationship with skill ratio (Figure 17(a)). However for an increasing number of experts in the combination, the relationship becomes much more nonlinear such that expert 1's weight is negative for any expert skill ratio that is much below 0.8 for any significant level of expert common correlation. Then in large combinations of forecasts ($\sim$ 10+) with expert correlations, even somewhat below

average performers will receive zero or a low negative weight. This is in contrast to a two-expert combination, where the expert needs to be substantially worse than the other expert to receive a negative weight.

From Equation 5.2, the point that the average weight of the crowd is 0 occurs when $Skr_1 = \frac{1}{\rho^2}$ and is not dependent on the number of crowd members (Figure 17(b)). As the degree of common correlation between experts increases, the weight placed on the better expert increases and the corresponding weights for the remaining crowd diminishes. This relationship is less about the weight of the crowd, as individual crowd members may still receive positive or negative weights, but instead this relationship highlights when expert 1 receives a weight of 1.0 or greater. In Figure 15 we saw that the weight of expert 1 converges to 1.0 at the same point for any k number of experts. This convergence occurs when the corresponding average weight of the crowd is 0. At the median value of average expert correlation observed in the economic data sampled, ~ 0.85, a skill ratio of only 1.38 is needed to cause the average optimal weight of the crowd to be 0.0 and the weight of expert 1 to be greater than 1.0.

Taken together the point where W_1 and W_{crowd} are zero defines a region (Figure 18) where expert 1 has a positive weight but not so positive as to drive the average crowd weight negative. Outside of this region for positive weights, the estimated weights are relying on the accuracy of the correlation estimation to optimize the combination by using negative weights. Although theoretically optimal, the literature on weight stability and non-negative weights suggest that negative weights should be avoided. It may be advantageous to drop the lower-skilled experts until the combination falls in this zone of positive weights. This region becomes quite narrow for common expert correlations above 0.8 which helps to explain the benefits observed of using a select crowd rather than all experts.



(a) Levels where expert 1's weight is 0 (b) levels where weight of crowd is 0

Fig. 17.: Skill ratio and degree of expert correlation where expert 1 and crowd's covariance optimal weights are zero (Equation 5.1 and 5.2).

5.2.3 Conclusions from expert correlations and estimated weights

Several key observations from this analysis can help guide a decision maker to make better forecast combinations:

- **Skill based weights** Even at lower levels of expert common correlation, correlation optimal weights will weigh the more skillful expert significantly more than variance based weights do, which is in turn more than equal weights. This suggest that even a low level of correlation will create a differentiated performance from variance based weights or from a simple average.

- **Skill based weights** With expert correlation and a higher number of experts, the most accurate expert could receive a weight approaching or greater than 1.0. Although theoretically optimal, it depends on a good estimate of the expert's skill. This situation would increase the impact of estimation errors on the overall combination.

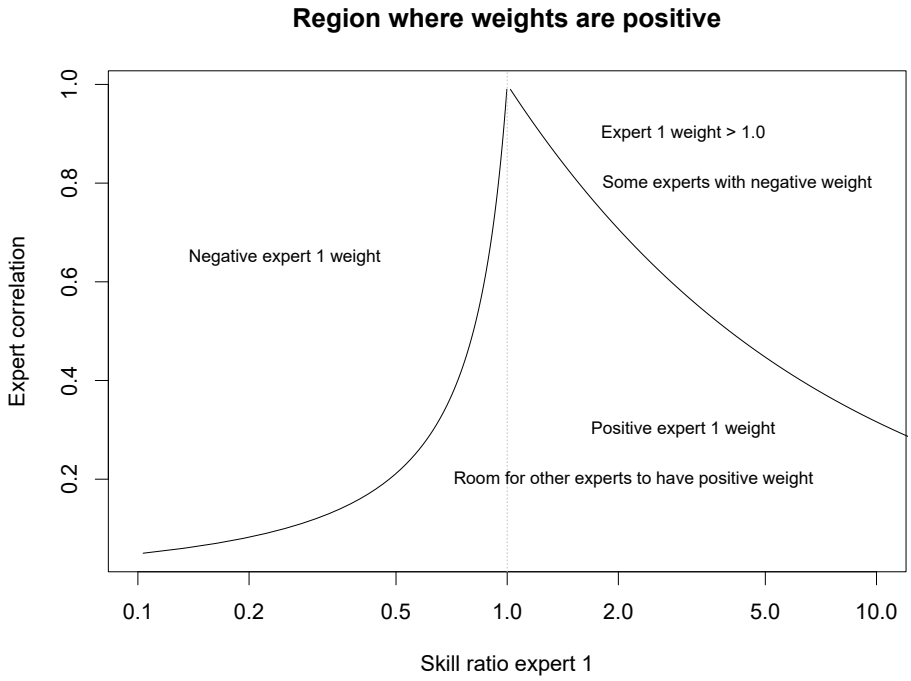


Fig. 18.: Region where common correlation optimal weights (Equations 5.1 and 5.2) are positive in a 10-expert combination.

- **Skill based weights** The assignment of weights to experts with expert correlation and a higher level of correlation is asymmetric with the better expert getting a larger positive weight than an expert who's poorly skilled to an equal magnitude.

- **Skill based inclusion** With large numbers of experts and a high degree of expert common correlation the weight of a slightly below average expert ($Skr \sim 0.8$) will be zero or negative.

- **Skill based inclusion** The number of experts and the common level of expert correlation in a combination, define a region of skill ratios where no expert has a weight less than zero or greater than 1. This region becomes quite narrow as

the level of common correlation increases above ~ 0.6.

These insights suggest that estimating weights at some level of correlation should improve the forecast combination, but erring on too high of a correlation level may create additional errors as individual weights become large. We will use these insights to develop revised and new decision rules for forecast combinations in the presence of expert correlation in Section 5.4. However, before embarking on a simulation, it will be important to understand the sensitivity of estimated weights (both variance optimal and the new common correlation optimal) to sample size when there is correlation between experts.

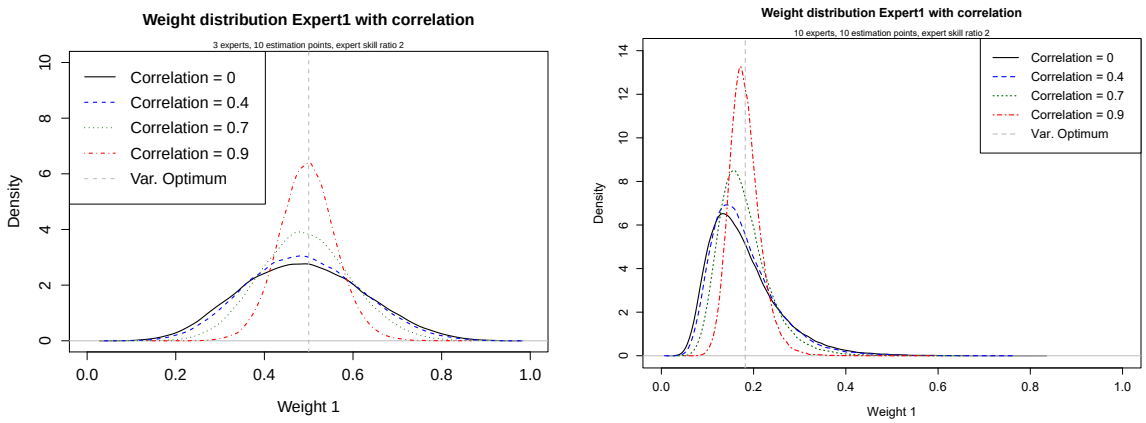
5.3 Estimation of weights with expert correlations

In Chapter 2 and 3 we saw that the estimated skill based weights can vary significantly from the true optimum. The shape and width of the estimated weight probability density curve was impacted by skill level, number of experts, and by number of estimation points. As a result, an expert cannot be 100% confident that an estimated weight will perform better than fixed weights in a simple average. We will now explore the distribution of estimated variance optimal weights and correlation optimal weights in the presence of various levels of expert correlation.

5.3.1 Estimation of variance based weights with expert correlation

Variance based weights are an attractive alternative to full covariance based weights as they only require estimation of k expert variances versus the full covariance matrix. Although, variance based weights do not try to estimate the level of expert correlation, it is likely that correlation is present and impacting the weight estimate. Figure 19 illustrates the impact of increasing expert correlation on the estimated weight distribution for a 3-expert combination and for a 10-expert combination where

expert 1 is twice as skillful as the other experts and all experts have the same level of correlation. These weights are based on the same simulation approach as previously described (Section 3.1) but with a sample from a multivariate normal distribution with the indicated level of common expert correlation. In both cases, increasing expert correlation narrows the distribution of the estimated weight and reduces the skewness in the distribution. This pattern of narrowing the distribution and reducing skewness persists in other combinations with differing levels of expert skill and number of experts.



(a) Three expert combination (b) Ten expert combination

Fig. 19.: Impact of expert correlation levels on estimated weight distribution for variance based weights with 10 historical estimation points based on a simulation. Expert 1 is twice as skillful as all other experts.

The overall impact of expert correlation on critical skill ratios for variance based weights is illustrated in Figure 20. The critical skill ratios have been estimated as before except that the simulation samples come from a multivariate normal distribution with the indicated level of common expert correlation. As the expert correlation increases, the decision maker can be more confident in their variance estimated weights.

This is a counter intuitive result, due to the effective degree of freedom in the estimated weights. As the correlation between experts increases, their forecasts are more constrained to be in the same direction and relative magnitude than with no correlation. Taken to an extreme, with near perfect correlation of errors, the expert forecasts will always be in the same rank order and have nearly the same relative spacing making it much easier to determine their relative rankings and corresponding weights. This insight suggests that expert covariance will actually make it easier to detect differences in expert skill. The reduced skill level needed as the overall expert correlation increases in Figure 19 supports this observation. Adjusting the critical skill ratio thresholds for expert correlation will enable a decision maker to benefit from this insight.

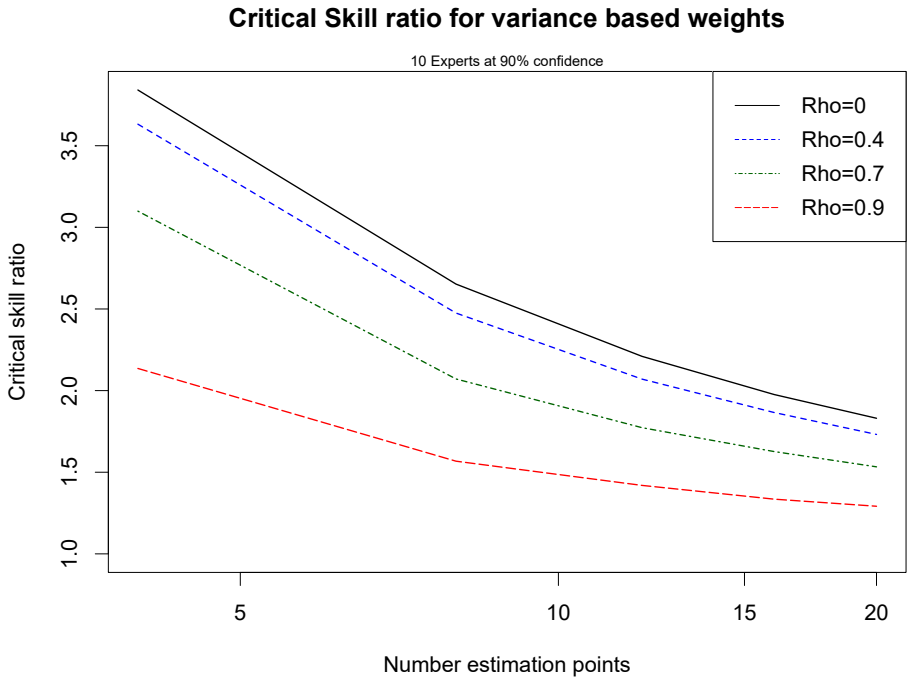
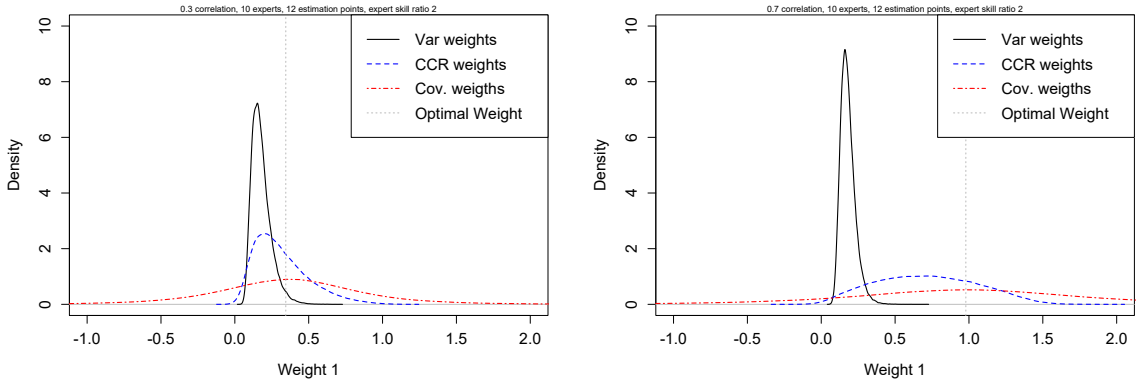


Fig. 20.: Impact of expert correlation levels on critical skill ratios for variance based weights at 90% confidence with 10 historical estimation points.

5.3.2 Estimation of Covariance based weights with expert correlation

A simplified way to account for expert correlation in the estimation of expert weights is to use one level of common expert correlation across all experts per the analysis in Section 5.2. In this case, we can use the average of all of the expert pair correlations in the estimation sample to determine the level of common correlation. We will call covariance optimal weights (Equation 2.1) that assume a common correlation structure: common correlation weights (CCR's) (see Appendix D). We can generate estimated weight distributions for both common correlation weights and for full covariance based weights (Cov) using a simulation similar to Section 3.1.1 but with sample points from a multivariate normal distribution with an inter-class correlation matrix at a known level of correlation.



(a) Expert error correlation 0.3 (b) Expert error correlation 0.7

Fig. 21.: Impact of expert correlation levels on variance, CCR, and covariance based weights for a combination with 10 experts, 12 estimation points and expert 1 has twice the skill of the crowd based on a simulation of 100,000 iterations.

Figure 21 illustrates the impact that correlation has on three approaches to estimating weights (Var, CCR, Cov). In both cases, covariance based weights have a

significantly wider distribution which is consistent with the many observations around the instability of covariance based weights (Section 5.1). The width of the Cov estimated weights is always greater than that of CCR, but they become much closer as the level of correlation increases. The CCR and Cov estimated weight distributions also become closer as the number of experts is reduced and they are equivalent when there are only two experts. Variance based weights are more narrow in width as they don't have the expert correlation estimation error. In fact as noted in Section 5.3.1, variance based weight distributions become more narrow as expert correlation increases. In summary we see that covariance weights have the highest variability followed by common correlation weights, and then variance based weights.

Although wider in distribution, the covariance estimated weights are more centered on the optimal weight (using Equation 2.1) than the CCR and the variance based weights. Variance based weights assume zero expert correlation and therefore do not enhance the weight of the higher skilled expert as covariance based weights will (Section 5.2) and are not expected to be close to the optimum weight. It is surprising to see that using an average value for the expert correlation induces some skewness in the estimated weight distribution as compared to covariance weights. For covariance weights, the estimation error in each expert pair correlation is weighted by the skill of the two experts (eq 2.1), this weighted average appears to have less bias. For CCR weights the estimation error in the average expert correlation is being replicated across all experts by design, giving it a greater effect. In summary we see that variance based weights have the highest estimation bias followed by common correlation weights and then covariance based weights with the lowest bias. Going forward, we will propose using common correlation weights as providing the error minimizing compromise between bias and variance estimation errors. We will test this recommendation in the simulation in Section 5.4.

The estimated weight simulation assumes an inter-class matrix so that there would be one known value of expert correlation ρ and a known optimal weight. This may be perceived as favoring CCR weights, however, the weight estimation process calculates each expert pair correlation separately from randomly drawn data so we believe that this simulation serves as a good illustration of the issues with covariance weight estimation (see Section 3.1.1 for the weight estimation process). The simulation in Section 5.4 will test the common correlation assumption by allowing the covariance structure be more random.

We can use the critical skill ratio estimation approach (Section 3.3.1) to determine the critical skill ratio where common correlation based weights are likely to be better than a simple average at a given level of confidence (Figure 22). Although increasing levels of correlation increases the width of the estimated distribution, the optimal value also moves significantly away from that of a simple average with the net result that a smaller sample size is needed to confidently choose a common correlation (CCR) based weight over a simple average.

5.3.3 Estimation of expert correlation

We have seen a large difference in the distribution of CCR estimated weights and covariance estimated weights. CCR weights assume that all experts have the same level of correlation. Estimating one common level of expert correlation should have less variability than estimating each individual expert pair correlation as there will be more data available. The decreased estimation variability may be sufficient to offset the weight estimation error induced by this common correlation assumption if the experts indeed have a similar level of correlation. Figure 23 illustrates the distribution of estimated expert correlation for an individual expert pair and for the average of all expert pairs. At a low level of expert correlation (Figure 23(a)),

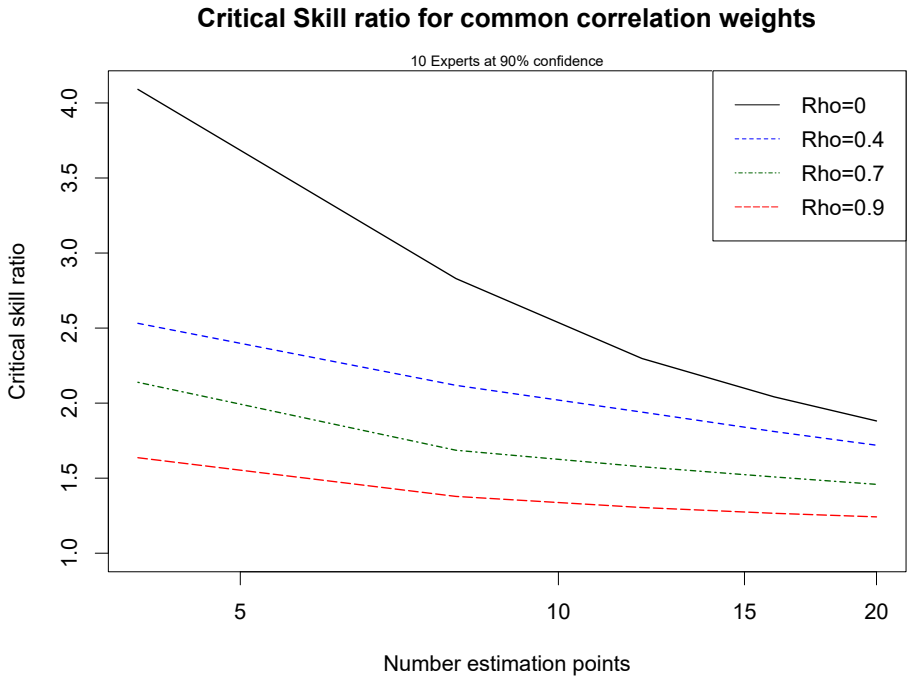
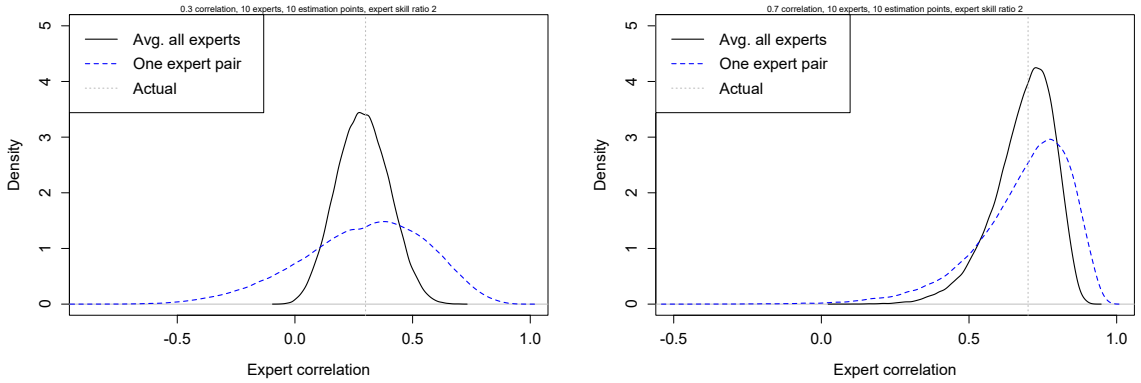


Fig. 22.: Impact of expert correlation levels on critical skill ratios for common correlation based weights at 90% confidence in a 10 expert combination.

the average correlation estimate has a much narrower distribution as expected. The distribution for one expert pair has wide tails and is likely to mis-estimate some extreme values. The likelihood of a mis-estimation will also increase with the number of expert pairs being estimated. In a 10-expert combination, there are 45 individual expert pair estimations. With a wide distribution, it is likely that several of these individual expert pair correlations will be large mis-estimations. Due to the sensitivity of weight estimation to the correlation level (Figure 13(b)), these estimation errors will be amplified in the estimated weights. The simulation in Section 5.4 will test this hypothesis and determine which is the larger effect: the error caused by assuming all experts have equal correlation versus the estimation error associated with estimating the correlation level of each expert pair.

It is interesting to note that the estimation benefit of an average level of expert correlation over that of one expert pair diminishes as the level of expert correlation increases. At a true correlation level of 0.7, the two estimation distributions become more similar (Figure 23(b)). In this case, because the additional data points used by an average estimate are highly correlated, their effective sample size and benefit are reduced. Although the benefit is reduced, the individual distribution is still more likely to mis-estimate an extreme value. The impact of mis-estimating a correlation is highly non-linear as the estimates approach 1.0, so even at higher correlation levels, there is an advantage to using a more conservative average estimate of expert correlation.



(a) Actual expert correlation = 0.3 (b) Actual expert correlation = 0.7

Fig. 23.: Comparison of the distributions for estimated expert correlations for only one expert pair to the distribution of the average of all expert pairs. The distributions are based on simulated values assuming an inter-class correlation matrix with 10 experts and 12 historical estimation points. Expert 1 skill is twice that of the other experts.

5.4 Improvements to forecast combination with expert correlation

In the analysis of the impact of expert error correlation on optimal skill based weights (Section 5.2 and 5.3), we observed that correlation significantly changes the optimal weight that should be given to an expert. However, using an estimated level of correlation in the calculation of the optimal weight also increases the potential for error in the estimated weight. These observations raise several questions that we will explore using a simulation approach:

1. **Weight estimation methods in the presence of expert correlations** Which of the four approaches discussed (SA, Var, Cov, and Common Correlation weights) provides the greatest accuracy improvement over an average expert while being robust when compared to a simple average?

2. **Correlation estimation methods** Which expert correlation estimation method will provide the greatest forecast combination accuracy improvement over an average expert while being robust when compared to a simple average?

3. **Decision criteria for using differential weights** When do the benefits of estimating an expert's weight with correlation out weigh the estimation errors?

4. **Decision criteria for when to differentially exclude experts** Can the estimation of expert correlation be used to improve the decision of which experts to exclude from a forecast combination?

The analysis of expert error correlation's impact on optimal skill based weights (Section 5.2 and 5.3) made two simplifying assumptions: a common correlation between all experts, and equal skill for members of the crowd. As we explore the above three questions, we will relax these two assumptions in the simulation by drawing the expert correlations and the skill of each crowd member randomly from distributions.

Taken together, these four simulation studies in the presence of random correlations between experts, will be used to develop a new decision rule for the combination of multiple experts in the presence of expert correlations.

5.4.1 Weight estimation methods in the presence of expert correlation

In Section 2.1 we introduced three common ways to estimate forecast combination weights. Then in section 5.3 we proposed a new, fourth approach called common correlation weights (CCR) which assumes one overall correlation level between every expert pair. The four weight estimation scenarios we will investigate are:

1. **Simple Average** In this method all experts are given the same weight regardless of their perceived skill or correlations with each other (Equation 2.5) . This approach avoids any estimation errors.

2. **Variance based weights** In this method experts are weighted according to their relative skill level (Equation 2.2) where skill level is measured as the inverse of the variance of their forecast errors. This method estimates each expert's skill but ignores their levels of correlation.

3. **Common correlation based weights** In this method experts are weighted according to their relative skill level and one level of common correlation which is assumed to be the same for all expert pairs (Equations 4.1 and 4.2). Similar to variance based weights this approach also using the same estimation of skill. However, unlike variance based weights, each expert's weight is adjusted for the common correlation. This has the effect of increasing the weight of the more skilled expert over a variance based weigh and vice versa (see Section 5.2). In this simulation we will use the average of the estimated correlation for each of the expert pairs in the forecast combination. In the next section we will explore

other estimators of the common correlation.

4. **Covariance based weights** In this method the experts are weighted by their relative skill and by the relative levels of expert correlations with other experts per equation 2.1. This approach requires estimation of each expert's skill and the level of correlation between each expert pair.

5.4.2 Common expert correlation estimation methods

We have seen that estimated weights are highly sensitive to the level of expert error correlation and to errors in the estimation of expert correlation. Four approaches to estimating expert error correlation will be investigated to determine the optimal error reducing tradeoff between the benefit of considering correlation in the estimated weight versus the risk of increased estimation errors. The scenarios to be investigated include:

1. **Estimation of average correlation across expert pairs** This is a revised approach where the variance of each expert is estimated separately from the correlation. In this case the historical data is pooled to estimate the average correlation across all of the expert pairs. The additional data points may improve the correlation estimate assuming that the estimation error is larger than the variation of expert correlation between expert pairs. A covariance matrix is then constructed using the estimates of each expert's variance and an inter-class correlation matrix with the estimated mean expert pair correlation. Weights are then estimated from the covariance matrix per Equation 2.1.

2. **Estimation of minimum expert correlation across expert pairs** Similar to the average expert correlation approach, this approach also separates the estimation of the expert variance from the correlation. In this case, the lowest,

positive expert pair correlation is used to form the inter-class correlation matrix. Zero is used in the event the lowest estimated expert pair correlation is negative. This non-negative correlation constraint is based on the assumption that expert correlations are due to shared data and models which create a positive correlation (Clemen, 1986; Elliott, 2011; Winkler, 1981). This approach recognizes the presence of expert correlation but may limit the impact of estimation error with a more conservative estimate of the expert correlation.

3. **Gird search for an optimal level of correlation** In this approach the experts skills are estimated from the historical estimation sample. Then a grid search is conducted in the historical estimation sample to find the level of common correlation which maximizes the frequency that the common correlation weights are better or equal in accuracy to a simple average. In the case of ties, the correlation level that also minimizes the mean absolute error in the estimation sample is chosen.

4. **Assume an exogenous level of expert correlation across expert pairs** In this case we are assuming the decision maker has some prior knowledge to set a correlation level. We will initially assume zero expert correlation which becomes variance weights (Eq. 2.2). This eliminates all estimation error but also forgoes any benefit of estimated weights with correlation. In addition, we will investigate scenarios where the level of correlation is assumed a priori to be 0.3 and 0.5 for all expert pairs. Based on observed data (Figure 12a) the median expert correlation level is never below 0.3 and frequently higher than 0.5. This observation of positive expert correlation levels has also been observed in the literature (Clemen, 1986; Elliott, 2011).

5.4.3 Decision criteria for using differential weights

Using a critical skill ratio to decide when to estimate weights was found to be an effective combination strategy in the simulations with no expert correlation (Section 4.3.2). This approach attempts to minimize estimation error by not estimating weights when the benefits are unlikely to outweigh the risks of estimation error. We will assess if the benefits of using a critical skill ratio persist when the experts are correlated. Two decision methods will be tested:

1. **Best90 with correlation weights** In this scenario, estimated weights are used for all experts if any one or more of the expert's skill ratio is either above or below the respective critical skill ratio thresholds estimated at a 90% level of confidence. In this case, the estimated weight will be the most error minimizing estimated correlation method.

2. **Select90 with correlation weights** In this scenario, estimated weights are used only for those experts whose skill ratios are outside of the 90% confidence critical skill ratio threshold (high or low). The unassigned weight will be averaged across the remaining experts within the thresholds. Again, the estimated weight will be the most error minimizing estimated correlation method.

These two approaches will be compared to the AIC approach proposed by Schmittlein et al. (1990) which uses an AIC statistic, estimated on the historical sample data, to choose between a simple average, variance based weights, and covariance based weights for the forecast period.

5.4.4 Decision criteria for when to differentially exclude experts

Work with expert clusters, LASSO regression, and with a select crowd of the top five experts has shown promising results for stabilizing expert weights and improving

the accuracy of a forecast combination. Highly correlated experts dilute the impact of more skillful experts while adding little additional information to the combination. Two potentially better methods for excluding experts will be explored:

1. **Drop90-Best90 with correlation weights** In this scenario, experts with skill ratios below the critic skill ratio at 90% confidence are dropped. Then estimated weights are used for all remaining experts if any one or more of the remaining expert's skill ratio is either above or below the respective critical skill ratio thresholds estimated at a 90% level of confidence. In this case, the estimated weight will be the common correlation method.

2. **Drop negative weights with correlation** In this scenario, experts who have a negative estimated weight using the common correlation weight approach are dropped. Then correlation based weights are recalculated for the remaining experts. This process is repeated iteratively until there are no negative weights. This approach is similar to a LASSO regression approach but with the important exception that the expert covariance structure is an assumed constant and is not estimated along with the parameters.

These two approaches will be compared to the select crowd approach that takes a simple average of the top five performing experts (Mannes et al., 2014). In addition we will include the Drop90Best50 approach which assumes no correlation (variance based weights) and performed well in the no correlation simulation.

5.4.5 Simulation methodology with expert correlations

We will use the same overall approach outlined in Section 4.3.1 to simulate each expert's skill and then estimate the relative performance of the forecast combination

rules to be tested. This approach will test the impact of varying skill levels across all experts versus just the one expert and the crowd. However, unlike the previous simulation which assumed no expert correlations, we will now also draw random levels of correlation between each expert pair in the combination.

In this simulation, we will assess performance of forecast combination approaches on two dimensions, performance and risk, as follows:

- **Performance** The performance of a combination approach will be measured as the average percentage improvement of the absolute error over an average expert across the simulation points. This performance will be compared to the same metric for a simple average to gauge performance relative to a simple average.

- **Risk** The frequency that the combination approach performs less accurately than a simple average.

Decision makers can play it safe and use a simple average, or they can use one of the other methods that, on average, performs better than a simple average but take the risk that this particular combination, due to the randomness of the estimation process, may turn out to be worse than a simple average. To be robust, the combination approach should display better overall performance and frequently be no worse than a simple average so that decision makers can be confident that they will have a better result than the benchmark of a simple average.

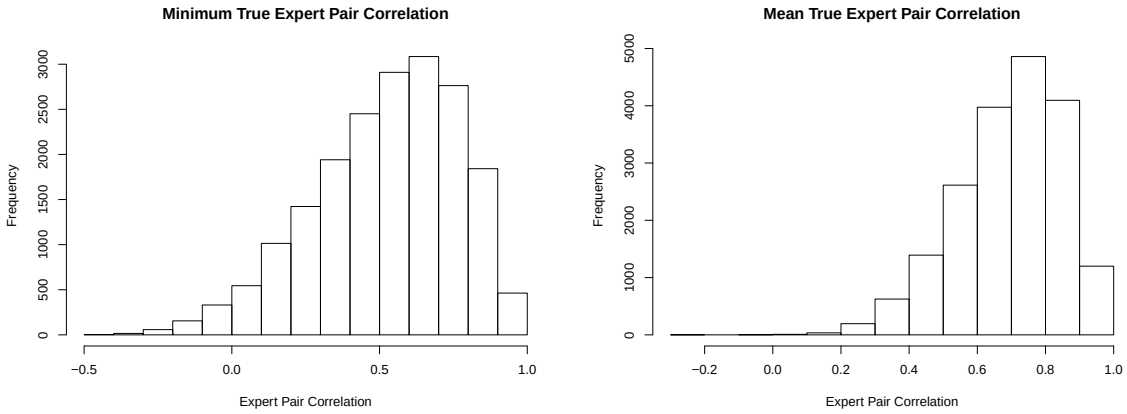
5.4.5.1 Expert pair correlation simulation

We can enhance the simulation approach used in Section 4.3.1 to include expert correlation by drawing the expert errors from a multivariate normal distribution with a zero mean vector and a random positive definite expert covariance matrix. To

generate a sample of the covariance matrix, we will first generate a random correlation matrix and then apply the individual expert skill, sampled as before, to create a random covariance matrix.

In order to use a multi-variate normal distribution, we must randomly sample expert pair combinations in a manner that results in a positive definite correlation matrix. To better represent expert correlations,we will only sample positive random values, as expert correlations are most frequently due to shared data, models, and assumptions which create positive correlations. Consistent with observed expert pair correlations from economic data (see Figure 12(a) and (b)), we will sample from a distribution that is heavily skewed towards values greater than 0.4. We are able to meet these three criteria by first sampling an overall expert correlation level ρ from a Beta(7,3) distribution. The overall level sampled is used to construct an inter-class correlation matrix where each expert pair has the same level of correlation. Inter-class correlation matrices are positive definite for values where $-1 < \rho < 1$. We then sample a random correlation matrix from a Wishart distribution with 28 degrees of freedom and centered on the inter-class matrix previously sampled. This process is repeated for each simulation point and provides a reasonable distribution of expert pair correlations as illustrated in Figure 24.

Once a correlation matrix with random expert pair correlations is created, it is used to generate a covariance matrix by multiplying either side by a diagonal matrix of the previously sampled expert MAE's multiplied by $\sqrt{\frac{\pi}{2}}$. The covariance matrix is then used to sample expert forecast errors from a multivariate normal distribution with a zero mean vector. The simulation then proceeds as described in Section 4.3.1.



(a) Lowest expert pair correlation (b) Average expert pair correlation

Fig. 24.: Sampling distribution of expert pair correlations used in simulation.

5.4.6 Simulation results

We will explore each of the four questions previously outlined in Section 5.4.0 using the above simulation approach.

5.4.6.1 Weight estimation methods in the presence of covariance

We looked at four approaches for estimating weights in a forecast combination with multiple experts and forecast correlations:

1. **Simple average (SA)**
2. **Variance based weights (Var)**
3. **Common correlation based weights (CCR)**
4. **Covariance based weights (Cov)**

The improvement performance of each approach is reported in Table 6 for combinations of 3, 10 and 28 experts and for 8 and 20 historical estimation points. Variance

based weights showed the greatest improvement over an average expert when there are only three experts in the combination. In this case the bias error of variance based weights is not as large as it becomes with increasing numbers of experts. With 10 experts in the combination variance based weights and common correlation based weights become comparable. A higher number of estimation points will give common correlation weights an edge over variance based weights as it is estimating and additional parameter, the average correlation level. With 28 experts common correlation weights has the highest accuracy improvement over an average expert. The additional experts provide more data with which to estimate the average correlation. Also, common correlation weights will provide greater differentiation between the experts (Figure 14) which becomes increasingly important as the number of experts increases. One might expect that with an increasing number of experts that the divergence in the correlation level between individual expert pairs would increase. However, the stronger performance of using an average correlation level with increasing number of experts suggest that this simplifying assumption is robust. Only with fewer experts (3 or 10) and 20 estimation points was a covariance weight able to beat a simple average. With 8 estimation points there is not enough sample to estimate the correlations between the many individual expert pairs. With 20 estimation points and 10 or less experts covariance is able to achieve a higher improvement than a simple average but this edge is lost when the number of experts increases to more than the number of estimation points.

In all cases, variance based weights have the lowest risk (Table 7) as measured by the frequency that combined forecasts have a higher absolute error than a simple average combination. Variance based weights by their construction will always be closer to a simple average than covariance based weights (Figure 14). Also they have the fewest parameters to estimate and therefore are the least variable. We will look at

	8 est. points			**20 est. points**		
Number of experts	3	10	28	3	10	28
SA	16.7	23.6	24.9	16.7	23.6	24.9
Var	**24.2**	**35.5**	37.8	**25.3**	36.2	38.3
CCR	14.6	34.4	**45.3**	19.1	**39**	**49.4**
Cov	8.2	-21.7	NA	19.4	25.6	1.2

Table 6.: Average percentage improvement (⇑ better) of a forecast combination absolute error over an average expert for various methods of estimating expert weights.

alternate correlation estimation methods, as well as, improving the decision of when to use an estimated weight versus a simple average as a way to improve on the risk performance of common correlation based weights.

	8 est. points			20 est. points		
Number of experts	3	10	28	3	10	28
Var	**38.2**	**29.5**	**24.4**	**36.5**	**27.2**	**21.4**
CCR	43.8	36.5	31.1	41.9	34.2	28.7
Cov	45.5	52.9	NA	41.9	38.2	56.6

Table 7.: Frequency that the forecast combination had a higher absolute error (⇓ better) than a simple average for various methods of estimating expert weights.

5.4.6.2 Correlation estimation methods

We looked at four approaches for estimating a common expert correlation level and how they impact the performance of common correlation estimated weights:

1. **Estimation of average correlation across expert pairs**
2. **Estimation of minimum expert correlation across expert pairs**
3. **Grid search for an optimal level of correlation**
4. **Assume an exogenous level of expert correlation across expert pairs**

The improvement performance of each approach is reported in Table 8 for combinations of 3, 10 and 28 experts and for 8 and 20 historical estimation points. It is interesting to note that the average level of expert correlation used in Tables 6 and 7 is actually the worse performing estimator of the common correlation level.

Both the minimum estimate of the correlation and the grid search to find an optimal value for the common correlation level consistently performed better than using the average value as the estimate. This suggests that a conservative estimate

of the correlation level is more optimum than an average estimate. Estimating too high of a level can increase the estimation error from the better skilled experts while also decreasing the weight of a lesser skilled , but perhaps more independent expert.

Overall, an assumed exogenous level of expert correlation of 0.3 had a lower error than all other methods (including a simple average) or, on two occasions, was within a few tenths of a percent of the best approach. Additional simulation runs show that this remains generally true for combinations with 3 or more experts and 8 or more estimation points. An assumed level of correlation of 0.3 performs better than a 0.0 assumption (variance based weights) as the weight versus skill curve (Figure 14) is steeper than the curve at a zero level of assumed correlation and will cause the estimated weights to put more weight on the more skillful expert and take a greater advantage of the estimated skill differences. Even so, the curve is not so steep as to greatly exaggerate estimation errors. Also in this scenario it is likely that most of the true expert pair correlations in the combination (Figures 12 and 24) are greater than the assumed correlation level of 0.3 so there is little danger of under-weighting pairs that are relatively independent but perhaps less skillful.

However there is a limit to the benefit of an assumed correlation. Increasing the assumed correlation to 0.5 slightly decreases performance versus a 0.3 assumption. In this case, the forecast combination is estimating weights on a steeper curve which will amplify true differences in skill as well as errors in skill estimation. Also in this case there will be more expert pairs where the assumed level of correlation may indeed be lower suggesting greater independence that would not receive an appropriate weight due to their relative independence.

The approaches of estimating one level of expert correlation (average or minimum) and then assuming that level for all expert pairs in the combination performed better than both a simple average and assuming a zero level of correlation (variance

	8 est. pts			20 est. points		
Number of experts	3	10	28	3	10	28
Simple average	16.7	23.6	24.9	16.7	23.6	24.9
Var. weights , $\rho = 0.0$	**24.2**	35.5	37.8	25.3	36.2	38.3
CCR, $\rho = average$	14.6	34.4	45.3	19.1	39	49.4
CCR, $\rho = minimum$	18.6	39.6	45.6	21.8	**42.9**	50.5
CCR, $\rho = gridsearch$	17.2	35.4	46.6	21.5	39.6	50.2
CCR, $\rho = 0.3$	23.8	**40.1**	**48.6**	**25.5**	42.7	**52.1**
CCR, $\rho = 0.5$	22	38.8	47.6	24.4	42.7	51.9

Table 8.: Average percentage improvement (⇑ better) of a forecast combination absolute error over an average expert for various approaches of estimating expert correlation ρ

based weights) when there were more experts in the combination. As there are more experts, there are more data points to improve the estimate of an overall correlation level. However using the minimum estimated expert pair correlation as the overall value consistently outperformed the average estimated level in all scenarios. This consistently higher performance level highlights the importance of a conservative estimation of a common expert correlation level which avoids under-weighting what may be relatively independent experts.

In all cases, except for a simple average, performance improved with increasing number of experts. This improvement was most pronounced when a common correlation (fixed or estimated) was used than with a simple average or with variance based weights (Rho=0). The difference in relative improvement is due in part to the larger role that correlation plays in estimated weights as the number of experts increased as

observed in Section 5.2. This difference may also be an artifact of the simulation, as with more experts, the random likelihood that a few experts are highly differentiated will increase.

The increase in the number of estimation points from 8 to 20 overall had a low but positive impact on the overall performance of each method. The methods that estimated the level of correlation benefited the most from the additional sample size, while the methods that assumed a level of correlation were less sensitive to estimation sample size.

The relative risk, as measured by the frequency an estimate is worse than a simple average, of each correlation estimation approach is summarized in table 9. It is interesting to note that the approach that used the minimum expert pair correlation (rounded to above zero) consistently has a lower risk than using the average estimate and similar to that of assumed fix correlation levels of 0.3 and 0.5. We see then that in both the average performance dimension and in the risk dimension, using a lower estimate of the common correlation more frequently minimizes the combination error. Similar to what was observed for performance, the risk of all combination approaches improved significantly with increasing number of experts as the improved performance was better able to offset estimation errors.

Although a better performer, using common correlation based weights also takes on a greater risk of being worse than a simple average. To be truly robust, we need an approach that both has a significantly higher performance than a simple average without taking on much additional risk of being less than a simple average. We will next investigate if using critical skill ratios to decide when to use a complex weight can help address this need when aggregating correlated experts. In the additional scenarios to be considered we will use common correlation weights with an exogenous assumption of a 0.3 expert correlation level. We will call these CCR3 weights.

	8 est. points			20 est. points		
Number of experts	3	10	28	3	10	28
Var, $\rho = 0.0$	**38.2**	**29.5**	**24.4**	**36.5**	**27.2**	**21.4**
CCR, $\rho = average$	43.8	36.5	31.1	41.9	34.2	28.7
CCR, $\rho = minimum$	42.4	32.2	26.3	40.7	30.4	24.2
CCR, $\rho = gridsearch$	42.8	36	30.1	40.7	33.2	27.4
CCR, $\rho = 0.3$	39.7	32.5	28.3	38	29.9	25
CCR, $\rho = 0.5$	41.2	34.6	29.9	39.4	31.9	26.9

Table 9.: Frequency that the forecast combination had a higher absolute error (⇓ better) than a simple average for various approaches of estimating expert correlation ρ.

5.4.6.3 Decision criteria for differential weights

We simulated three approaches for deciding when to estimate weights in the presence of expert correlations.

1. **Best90 / Best98** Estimate weights for all experts when one or more expert is beyond the critical skill ratio (high or low) at the stated level of confidence (90 or 98%). Estimated weights will assume a fixed common correlation level of 0.3.

2. **Select90 / Select98** Estimate weights only for experts who are outside of the critical skill ratios at the stated level of confidence. Estimated weights will assume a fixed common correlation level of 0.3.

3. **AIC** Choose the combination approach (simple average, variance based, or covariance based weights) based on the lowest AIC statistic for each approach

as evaluated on the historical sample data. The covariance based weights are determined by estimating the correlation of each expert pair.

The improvement performance of each approach is reported in Table 10 for combinations of 3, 10 and 28 experts and for 8 and 20 historical estimation points. Overall, all three approaches consistently out performed a simple average demonstrating the advantage of a decision based combination approach. Similar to the results in the simulation with no covariance (Table 4), the Select90/98 approaches performed no better than the simpler Best90,98 approaches. The estimation errors avoided by the Select approaches are low as the experts within the critical skill ratio thresholds all have similar skills and their estimated weights will be close to using an average. However, the chance that some experts do not receive a differentiated weight when they should, lowers the overall performance of the Select approach. Both the Best90 and Best98 approaches, using an assumed expert correlation of 0.3, consistently outperform the AIC approach. The difference in performance is small when there are only three experts but becomes sizable with 10 or more experts. This improvement is primarily due to the strong performance of using an assumed correlation level when estimating weights.

At lower historical sample sizes the performance of the Best approaches are on par with that of always using a correlation based weight. The Best approach will avoid some of the higher risk of an estimation error due to fewer data points, but may also miss an opportunity to differentially weight the experts. As the number or estimation points increases, the Best approaches of choosing when to use an estimated weight lag the performance of always estimating weights. In this case, the additional data points increase the accuracy of the estimated weights. This loss of performance, however, from using a Best approach, is mitigated somewhat by the additional data

points that help to make a better decision. The differences in performance between the Best90 and Best98 are marginal. We will need to look at the relative risk of each method to see if the small loss in performance by using a Best decision approach is outweighed by a reduction in the risk of a bad estimate.

The AIC method of choosing which weighting method to use defaults to always choosing a simple average when there are only 8 estimation points and 10 or more experts. Similarly with 20 estimation points it defaults to a simple average with 28 estimation points. In these cases the penalty for the larger number of parameters in an estimated weight swamps the gain in estimated likelihood.

	8 est. points			**20 est. points**		
Number of experts	3	10	28	3	10	28
Simple average	16.7	23.6	24.9	16.7	23.6	24.9
AIC chooses weights	21	23.6	24.9	23.6	29.2	24.9
CCR $\rho = 0.3$	23.8	**40.1**	**48.6**	**25.5**	**42.7**	**52.1**
Best90 with CCR $\rho = 0.3$	23.9	39.9	48.2	**25.5**	42.6	51.9
Select90 with CCR $\rho = 0.3$	**24**	39.3	47.2	**25.5**	42.1	50.6
Best98 with CCR $\rho = 0.3$	**24**	40	48.2	**25.5**	42.1	50.5
Select98 with CCR $\rho = 0.3$	23.9	37.7	44.2	**25.5**	41.4	49.1

Table 10.: Average percentage improvement (⇑ better) of a forecast combination absolute error over an average expert for various approaches of deciding when to use an estimated weight versus a simple average.

The risk for each decision method as measured by the frequency of performing worse than a simple average is reported in Table 11. Overall AIC consistently has the lowest risk as it most frequently chooses to use a simple average. This low risk

also corresponds to the smallest improvement in performance. In the most risky scenario, with only 3 experts and 8 historical estimation points, AIC has a similar level of performance as other methods with much less risk. However, as the number of experts increases, the performance of the Best approach improves substantially over AIC while their risks remain flat to decreasing, making them a better choice in these situations.

Using Best90 or the Best98 approaches with common correlation based weights, significantly reduces the risk of using estimated weights with little loss in performance. Best98 sets a higher bar for when to select an estimated weight than Best90, yet has a minimal loss in performance between the two approaches. Best98 chooses to use an estimated weight $49-70\%$ of the time while Best90 chooses to use an estimated weight $71-97\%$ of the time. By choosing not to estimate weights as frequently, Best98, avoids the occasional large estimation error while capturing the bulk of the improvement benefit.

Overall using a critical skill ratio approach, in this case Best98, reduces the risk of using an estimated weight by half with only a $1-2\%$ reduction in performance. Best98 offers a $1.5-2$X improvement in performance over a simple average with only a $\sim 15\%$ risk of being worse than a simple average. This improvement in performance at a low risk makes the Best98 a robust approach for using estimated weights with an assumed 0.3 level of expert correlation.

	8 est. points			20 est. points		
Experts	3	10	28	3	10	28
AIC chooses weights	**8**	**0**	**0**	**12.7**	**2**	**0**
CCR $\rho = 0.3$	39.7	32.5	28.3	38	29.9	25
Best90 with CCR $\rho = 0.3$	25.7	25	24.8	27.1	26	24.2
Select90 with CCR $\rho = 0.3$	25.7	25.3	25.3	27.2	26.7	26.1
Best98 with CCR $\rho = 0.3$	15.8	16.1	15.8	19.8	14.6	13.1
Select98 with CCR $\rho = 0.3$	15.8	16.7	16	19.8	14.9	13.5

Table 11.: Frequency that the forecast combination had a higher absolute error (⇓ better) than a simple average for various approaches of deciding when to use estimated weights.

5.4.6.4 Decision criteria for differential exclusion

Assuming a fixed, common level of expert correlation coupled with using estimated weights when there is sufficient expert skill ratio, has been shown to be a robust way to combine experts. We will now examine the impact of excluding those experts with lower skill who may be diluting the responses of the more skillful experts as a third approach to improve upon a simple average. Three new approaches for selecting which experts to exclude were simulated and compared to other benchmark combination strategies.

1. **Drop90 Best50 with R=0.0** This is an approach that performed well in the no correlation simulation (Section 4.3.4) and continues to assume no correlation between experts. In this approach, experts with skill level less than the lower 90% critical skill level are dropped. Then the remaining experts are combined with variance estimated weights (R=0.0) if the skills of the remaining experts

are sufficiently dispersed as tested by exceeding a critical skill ratio threshold of 50%. If not sufficiently dispersed, then a simple average is used.

2. **Drop90 Best50 with R=0.3** This is an extension of the previous approach that leverages the learnings on correlation estimation and assumes a fixed, common expert correlation of 0.3 when estimating critical skill ratios and weights.

3. **Drop experts with negative weights with R=0.3** This approach estimates weights using a common expert correlation level of 0.3. However, any expert with a negative weight is then dropped, and the weights are re-estimated for the remaining experts. This process continues iteratively until there are no negative weights. This is a different way to achieve the benefits seen in the literature from constraining the weights to be positive.

4. **Drop experts with negative weights, then use the Best98 approach** This approach is an extension of the previous approach, where after all positive weights have been achieved, the Best98 criteria is used to determine if there is enough dispersion in the remaining experts' skill to justify using an estimated weight. If the expert skills are not significantly differentiated, a simple average is used, similar to the Best approach in Section 5.4.2.2.

In addition to the above new approaches, we will include for comparison purposes the select crowd approach from Mannes et al. (2014) as well as a simple average, the estimated weights with a common correlation of 0.3 and the Best98 approach from the last two simulations.

The improvement performance of each approach is reported in Table 12. None of the approaches that exclude experts performs better than the Best98 approach with an assumed expert correlation of 0.3. This suggests that the enhanced weighting of the

more skillful experts that an assumed level of expert correlation provides, is sufficient to mitigate dilution effects of more experts. This is not surprising given the slope of the weight versus skill level line in Figure 13(b). This result also recognizes that each expert, even when highly correlated, has some unique information to contribute and therefore should be included with an appropriate weight.

The approaches that drop experts below a certain skill level consistently performed slightly below that of the Top5 approach. As seen in Figure 25 the Drop90 approach often retains many more experts than the Top5 approach. In a separate study, reducing the confidence level to Drop50 did not improve this result as a much more radical cut is needed to achieve the reduction. The strategy of including the few, most accurate experts, versus dropping the many worse experts, appears to be a better approach.

The approaches that dropped experts with negative weights performed consistently, but marginally better than the Top5 approach. As seen in Figure 25 the number of experts retained distributions are bimodal. This is due to the 50/50 combination of highly dispersed and not highly dispersed skill level populations in our simulation. The peaks in the lower number of experts retained are associated with the more highly dispersed skill level population in the simulation, while a larger number of experts were retained when the simulation was sampling from a less dispersed skill level distribution. When the skill level dispersion was high, the drop negative weights approach automatically retained 5-7 experts out of 28 possible experts, which is very similar to the recommendations of the select crowd approach (Mannes et al., 2014). When the skill levels are less dispersed, the drop negative weights selects most of the experts which is also similar to the recommendation of the select crowd approach. The key difference between these two approaches is that the Top5 approach uses a simple average after experts have been excluded while the drop negative weights uses

estimated weights with an assumed common expert correlation level. It appears that the use of this weight estimation method helps give the drop negative weights approach a performance edge over the Top5 approach. A possible future improvement may be to try a Top5 approach with a Best98 weight estimation for the remaining Top5 experts.

	8 est. points			**20 est. points**		
Number of experts	3	10	28	3	10	28
Simple average	16.8	23.3	24.9	16.8	23.3	24.9
CCR $\rho = 0.3$	23.6	**40.1**	**48.6**	**25.5**	**42.7**	**52.1**
Best98 with CCR $\rho = 0.3$	**23.9**	40	48.1	25.4	42	50.5
Top5 with simple average	NA	32.6	42.2	NA	33.7	44.3
Drop90-Best50 with Var $\rho = 0.0$	23.7	37.8	41	24.8	39.5	42.8
Drop90-Best50 with CCR $\rho = 0.3$	23.1	38.5	45.5	24.4	40.7	47.8
Drop negative weights with CCR $\rho = 0.3$	23.5	38.4	44.1	25.3	40.7	46.8
Drop negative weights then Best98 with CCR $\rho = 0.3$	23.5	38	43.2	25.3	40.1	45.6

Table 12.: Average percentage improvement (⇑ better) of a forecast combination absolute error over an average expert for various approaches of deciding when to exclude an expert from the combination.

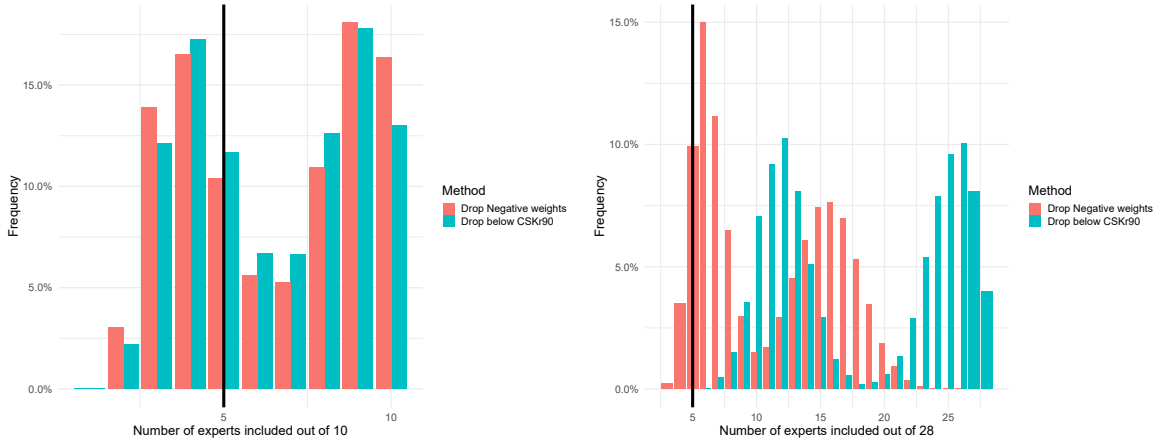
As seen in Table 13, none of the exclusion approaches, by themselves, significantly changed the risk of an estimate being worse than that of a simple average. This observation highlights the fundamental risk of choosing the wrong experts based on noise in the historical assessment data. It is surprising that the Best98 approach to mitigating risk is not nearly as effective after experts have been excluded than it was when applied to combinations of all experts. The Best98 approach is focused on reducing weight estimation error, but with the prior exclusion of less skilled experts,

the opportunity for estimation error is already reduced. The result is that Best98 has less impact.

Overall, the Best98 approach with an assumed correlation of 0.3 applied to all experts, also has both better performance and lower risk than any of the expert exclusion methods.

	8 est. points			**20 est. points**		
Number of experts	3	10	28	3	10	28
CCR $\rho = 0.3$	39.9	32	28.3	38.2	29.7	25.1
Best98 with CCR $\rho = 0.3$	**15.4**	**15.7**	**15.8**	**19.9**	**14.3**	**13.1**
Top5 with simple average	NA	34.2	29.3	NA	32.4	26.7
Drop90-Best50 with Var $\rho = 0.0$	39.4	29.8	24.9	38.3	28	22.5
Drop90-Best50 with CCR $\rho = 0.3$	40	31.8	28.6	38.8	29.3	24.9
Drop negative weights with CCR $\rho = 0.3$	39.8	31.7	28	38.1	29.3	24.4
Drop negative weights then Best98 with CCR $\rho = 0.3$	20.7	29.8	27.3	29.6	24.1	24.4

Table 13.: Frequency that the forecast combination had a higher absolute error (⇓ better) than a simple average for various approaches of deciding when to exclude an expert from the combination.



(a) 10 expert combination (b) 28 expert combination

Fig. 25.: Number of experts included in forecast combination after dropping those experts outside the 90% critical skill ratio or those with negative weights. Based on a simulation with 20 historical estimation points.

5.5 Discussion

The theoretical analysis in Section 5.2 shows that with expert correlations, significantly more weight should be put on the most skillful experts (Figures 13 and 14) even to the point of giving negative weights to the less skillful experts. As the independence between experts decreases, the less skillful experts bring less new information to the combination, however they also bring more noise. This effect puts a premium on the more accurate information provided by the more skillful experts. With many experts in a combination, this dynamic can frequently result in a few highly skilled experts with significant positive weights and many slightly negative weights for the less skillful experts.

The benefit from recognizing and using the degree of expert correlation in estimating weights causes the critical skill ratio threshold for when to estimate weights to shift towards preferentially estimating weights at lower skill ratios (Figure 22). In

the presence of correlation, there is a greater benefit for more accurate information, such that estimation errors can be overcome at lower skill ratios if the degree of correlation is known. This suggests a strategy of emphasizing the differential in expert weights in the presence of correlation to overcome the estimation errors as opposed to approaches that shrink the weights towards the value of a simple average.

The challenge to using the above insights is that estimating the level of expert correlation in a small sample is very difficult. Figure 23 illustrates the wide variation in estimated expert correlations. Figure 21 shows how attempting to estimate expert correlation increases the variability in the estimated weight. Multiple approaches to estimating expert skill, including making an exogenous assumption and avoiding an estimation step, were evaluated by a simulation.

Through simulation we have looked at three questions that are fundamental to developing a combined forecast from multiple expert forecasts that have correlated forecast errors. The simulations extend the theoretical discussion in Section 5.2 to situations where the experts do not all share the same level of correlation. We have found that:

1. **Best way to estimate expert correlation** The past problems of estimating a full covariance matrix can be resolved by using an inter-class covariance matrix that assumes a constant level of correlation between experts. The level of common correlation can be determined a priori based on decision maker experience or by using the lowest, non zero value from the estimate of each expert pair's correlation level. This approach significantly improves the performance of an expert combination above that of a simple average. With this improvement, what has been theoretically the optimal error minimizing approach for combining experts, also becomes the best approach in practice.

2. **Decision criteria for using differential weights** Using a critical skill ratio approach estimated at a 98% level of confidence to decide when to estimate weights, greatly reduces the risk of the combination performing worse than a simple average. This approach makes the common correlation weight estimation robust.

3. **Decision criteria for differential exclusion** Although a promising way to improve upon a simple average, exclusion methods do not perform as well, nor are as robust, as using all experts with common correlated weights and a CSKr98 decision threshold for applying estimated weights.

In summary, using weights with an assumed common correlation of 0.3 and with decision criteria of CSKr98 for the use of estimated weights provides a robust improvement over a simple average. We will now test this conclusion on economic forecast data and on experimental survey data.

CHAPTER 6

EMPIRICAL EXPLORATION OF MULTIPLE COMBINATION APPROACHES

The original question in this dissertation is "Can anything beat a simple average"? Although many theoretical approaches for differentially weighting experts have been developed, they do not consistently provide a better result than a simple average when applied to real data (see Sections 1.1, 4.1.1, and 5.1). Approaches to differential inclusion have been more promising where the Select Crowd approach (Mannes et al., 2014) has shown improvement over a simple average. We propose a third, new method for differential weighting which in a simulation study (Section 5.4) outperformed a simple average and the Select Crowd approach. In this Chapter we will assess these three approaches to forecast combination with real data to test the validity of our proposal.

The three approaches:

1. **SA** A simple average of all available forecasts.

2. **Top5** Average the five most skillful experts (Mannes et al., 2014). The level of skill is determined by the mean absolute error of the experts evaluated over the historical skill estimation period. We will extend this approach to fewer then five experts by using a simple average when there are five or fewer experts.

3. **CCR3 and CCR3.B98** In these approaches, an exogenous common level of expert correlation of 0.3 is assumed and then used to calculate covariance weights. In the case of CCR3.B98, estimated weights are only used if the skill level of

one expert is beyond the critical skill ratio thresholds (high/low) for common covariance weights evaluated at 98% confidence. Otherwise a simple average is used.

This is the first study, that the author is aware of, which directly compares the performance of differential weighting and differential inclusion approaches with real data. The results of this study will provide decision makers guidance on which approach most frequently has the lowest forecast combination error in practice.

6.1 Empirical data and methodology

Following the approach used by Mannes et al. (2014), we will run the assessment on two bodies of data, economic forecasts and "expert" estimations of known quantities by college students in various psychology experiments. The economic data will test the efficacy of each combination approach when the forecasts are made periodically for the same economic quantity over an extended period of time. In this case, the skill of the expert is determined by their historic forecasting performance of that particular quantity. The experimental data set will test the efficacy of each combination approach when a collection of "experts", in this case college students participating in the experiment, estimate multiple quantities all in the same domain of knowledge. For example, in what year did each of the following 20 authors win the noble prize for literature, or how much wealth have each of these 20 people from the Forbes list of the 50 wealthiest people been reported to have (Larrick et al., 2007)? The experimental data provides multiple questions in the same knowledge domain that can be used to assess the skill of each participant. The economic and experimental data each have different degrees of forecast variability and bracketing and different levels of forecast error covariance. Taken together, these different data sets will provide two unique tests of the four combination approaches.

6.1.1 Economic data

Economic data was furnished by a private economic consulting firm under a research data sharing agreement. Two sets of quarterly one-step-ahead forecasts and corresponding realizations were provided: one from G7 countries from 1Q2010 to 2Q2018, and one from a collection of emerging market countries from 2Q2000 to 3Q2018. These selections were determined by data availability and the desire for a more diverse set of countries than has been used in past economic forecast combination studies. The countries included were:

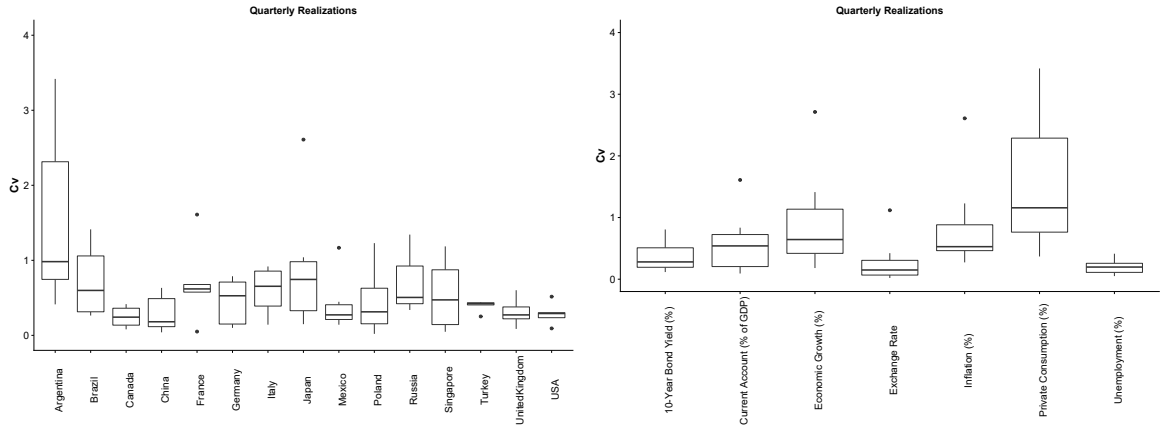
Emerging Markets	**G7**
Argentina	Canada
Brazil	France
China	Germany
Mexico	Italy
Poland	Japan
Russia	United Kingdom
Singapore	USA
Turkey	

The economic indicators forecasted were:

- 10-Year Bond Yield (%, eop)
- Current Account Balance (% of GDP)
- Economic Growth (GDP, ann. var. %)
- Exchange Rate (local currency per USD, eop)

- Inflation (CPI, ann. var. %)
- Private Consumption (ann. var. %)
- Unemployment (% of active population, aop)

In the case of economic indicators, the quantity being forecast changes over time. Indicators with greater temporal variations will make forecasting and skill estimation more challenging. The relative variability of the economic quantities being forecast are illustrated in Figure 26. The coefficient of variation for the realizations vary from a high of 7 (economic growth rate in Italy) to a low of 0.02 (exchange rate in Poland) with a median value of 0.42 across all indicators and countries. Economic data from the USA, which is frequently used for forecast combination studies, has relatively low variability during the time period sampled. Including the other countries will test the aggregation of the forecasts on a wider variety of conditions.

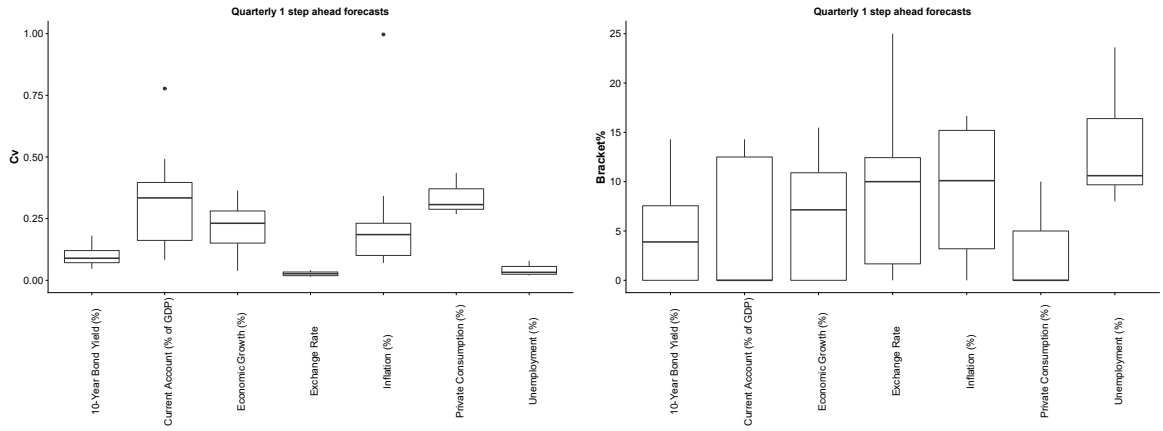


(a) By country (b) By indicator

Fig. 26.: Variability in economic indicator quarterly realizations as measured by the average over time of the coefficients of variation (Sd/Mean) for each series.

The dispersion in the economic forecasts around a mean forecast value and the degree to which it brackets the true value is illustrated in Figure 27. It will be more

difficult to differentiate expert skill and it will be less beneficial to differentiate expert skill in forecasts with low dispersion as the resulting combination will be close to that of a simple average. The economic data has some indicators with very low dispersion as well as three outliers with higher dispersions (Argentina - Current Account% of GDP 0.78 Cv, Argentina - Private Consumption % 0.43 Cv, Italy - Current Account % of GDP 0.46 Cv). All of the economic data has average bracketing levels below 25%. This indicates that most of the forecasts are frequently on either side of the true value. This situation will favor methods that can more heavily weight the more accurate forecasters as a simple average and will not benefit from offsetting forecast errors that are above and below the true value.



(a) Forecast Cv (b) Forecast bracketing

Fig. 27.: Quarterly one-step-ahead forecast variability and bracketing box plot of averages over time for each country

The level of forecast error correlations between expert pairs for the economic data is plotted in Figure 12. The overall median values of expert pair error correlations are above 0.3 as previously observed, however there are some forecast series with negatively correlated expert pairs (Figure12(b)). Of the five series with some negative

expert correlations, three are for the Inflation economic indicator. This represents 21% of the countries with inflation data reported. The negative correlations will favor the approaches that use an average (SA, Top5).

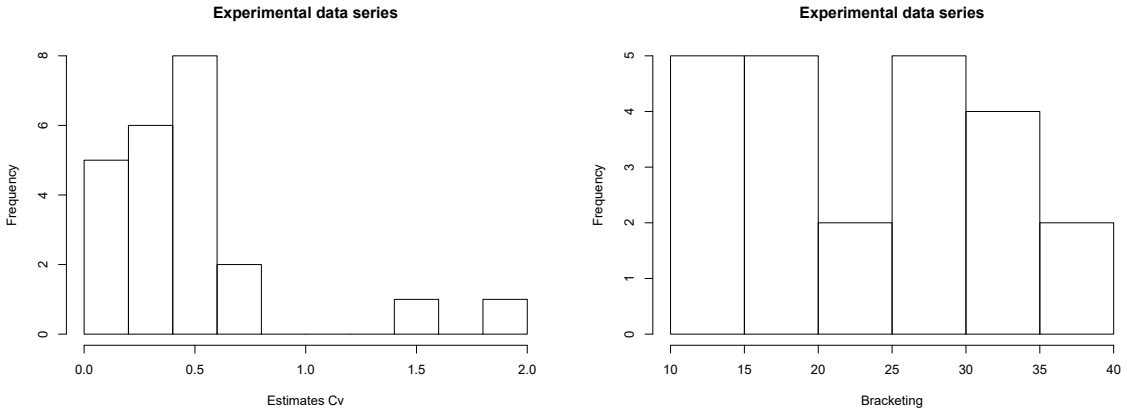
Forecast combinations were calculated using each of the four combination methods with historical sample sizes of $4, 8, 12, 16, 20$ to assess expert skill. For each indicator, in each country, one-quarter-ahead forecasts were selected where there were at least three experts reporting the required historical estimation sample size in the previous contiguous quarters. Experts with historical gaps in forecast reporting were not considered for estimated weights as there would not be a consistent way to estimate their skill. Then for each forecast period, the forecasts were aggregated using weights for each method calculated on the historical sample. The aggregated forecasts were then used to calculate an overall MAPE for each combination method for each country-indicator. MAPE was chosen as our error metric as it will normalize for differing levels between the many data series being analyzed. Only country-indicator pairs with more than 10 quarters of combined forecasts were reported. In total, for the 4-quarter estimation window, there were 74 unique country-indicators reported with a total of 2261 unique forecast aggregation points. As the historical estimation window increased to 20 quarters, the number of country-indicators with at least 10 quarters of aggregated forecasts was reduced to 26 with 541 unique forecast aggregation points. The results from this work will be discussed in Section 6.2.

6.1.2 Experimental data

Data compiled from three psychology studies where students were asked to estimate various quantities was provided by the original authors (Larrick et al., 2007; Soll and Larrick, 2009; Soll and Mannes, 2011). The students were asked to make their best estimates of numerical quantities in domains of general knowledge and news sto-

ries under controlled conditions. Only experiments with 20 or more questions in each domain where used so that we could evaluate the impact of various skill estimation sample sizes. In total, there were 23 series of knowledge domains with 20 or more questions asked to the same sample of participants. These series of domain specific questions had from 15 to 132 participants providing estimates, with the bulk of the studies having 15 to 20 participants.

The dispersion of the estimates and degree of bracketing is illustrated in Figure 28. The dispersion of the estimates around their mean value was considerably higher than that of the economic data with a median average coefficient of variation of 0.41. This should provide a better opportunity for combination approaches with differential weighting or inclusion to identify the better estimators. The median level of 25% that the estimated values bracketed the true value, was also considerably higher than the economic data. The higher degree of bracketing should favor approaches that use an average (SA, Top5).

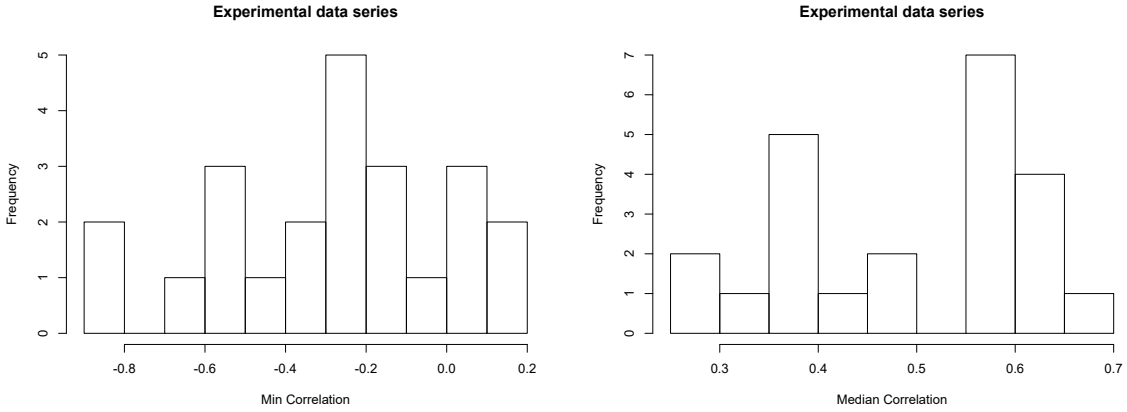


(a) Cv of estimates (b) Percent bracketing of estimates

Fig. 28.: Dispersion and bracketing of estimates

The level of error correlations between pairs of estimators for the experimental

data is plotted in Figure 29. Negative correlations between estimators was much more prevalent than the economic data with 78% of the series having some negatively correlated errors. In some cases, the negative correlation was as as large as -0.8. The median level of estimator correlation was always positive, but it is more frequently in the $0.4 - 0.6$ range versus the economic data which was more frequently in the $0.8 - 0.95$ range. It is not surprising that student estimators are less correlated as they are less likely to be using (or paying attention to) the same news outlets as professional economic forecasters following their respective markets. The lower levels of correlation generally, and the negative correlations specifically, will favor approaches that use an average.



(a) Minimum expert correlation (b) Median expert correlation

Fig. 29.: Correlation of participant errors by Experimental data series

For each of the questions in the 23 series, a random sample of $n \in (4, 8, 12, 16)$ other questions was sampled to estimate the skill of each participant. The estimates for the questions in each knowledge domain were asked contemporaneously of each participant so there is no logical order for determining the skill estimation sample. Weights and aggregate forecasts were then calculated for each question in the study

using the four combination methods. An overall MAPE was then calculated for each combination method for each series of questions. This entire process was then repeated 20 times for each series to reduce the potential variability from the random skill estimation sample. In total, there were 460 evaluations of series MAPE and 9600 unique instances of aggregated questions. The results from this work will be discussed in the next section.

6.2 Empirical exploration results

We will look at the empirical exploration results from several perspectives. First, we will look at how often a method beats a simple average as this is our primary focus. Then we will look at which method has the highest frequency of beating a simple average out of the three methods to understand which may be the preferred method. In addition, we will look at how well the methods perform on a risk perspective by considering how frequently they may place at a lower level of relative performance versus their peers. Finally, we will use a sign test to determine if the method proposed by this work is significantly better statistically than a simple average as well as each of the other methods..

6.2.1 Performance versus a simple average

For economic forecasts, the frequency that each combination approach had a better MAPE than a simple average is provided in Table 14. Similar to the simulation results, the proposed common correlation approach (CCR3) is consistently better than a simple average. However, 16 or more points are required for CCR3 to be differentiated from the Top5 approach on economic data. CCR3.B98 is the worse performer with 12 or fewer historical estimation points. However CCR3.B98 actually performs better than CCR3 with 20 points which is surprising as it was always close, but never better in the simulation. The additional historical estimation points enables CCR3.B98 to better resolve the threshold between when to use a CCR weight and when to use a simple average.

The application of the four combination approaches to the experimental data tells a somewhat different story (Table 14). In this case, CCR3 is still consistently better than a simple average as seen in the economic data and simulation. However,

	Frequency(%) MAPE better than SA								
	Economic forecasts					Experimental data			
Est. window	4 pts	8 pts	12 pts	16 pts	20 pts	4 pts	8 pts	12 pts	16 pts
Top5	56.8	63.2	56.9	61.4	53.8	**62.6**	**67.6**	**70.7**	**75.0**
CCR3	**60.8**	**64.7**	**56.9**	**70.5**	61.5	53.3	61.7	64.1	69.8
CCR3.B98	51.4	50.0	48.3	68.2	**69.2**	56.1	61.5	68.3	74.1
Sample size	74	68	58	44	26	460	460	460	460

Table 14.: The frequency that the MAPE of a forecast combination applied to each of the data series is less than the MAPE of a simple average (⇑ better) of the available forecasts.

Top5 and CCR3.B98 more frequently beats a simple average than CCR3. Both of these methods use a simple average in some form and therefore will benefit from the generally lower correlation levels and specifically from the negative correlations in this data set. This scenario illustrates the advantage of adding the Best98 decision to the CCR3 weighting process as it avoids using estimated weights when a simple average is most likely to work well. To avoid confounding our analysis with two effects (CCR3 and B98) we will focus our analysis on CCR3 going forward.

6.2.2 Best performance across all methods

We have seen that both CCR3 and Top5 can beat a simple average. Table 15 indicates which method is most frequently the most accurate of all methods. The newly proposed method, CCR3, performs best in all but one scenario with the best improvement achieved when there are 16 or more estimation points. A chi square test confirms that there is differentiated performance in each scenario except for the economic data with 12 estimation points and experimental data with only 4 estimation

points. However, this test does not imply that any one method is better than another. We will look at a paired comparison in Section 6.2.4. The CCR3 method does best on the economic data where there is more positive expert error correlation. However, there is sufficient correlation even with student estimators to provide a performance edge for CCR3 in the experimental data set.

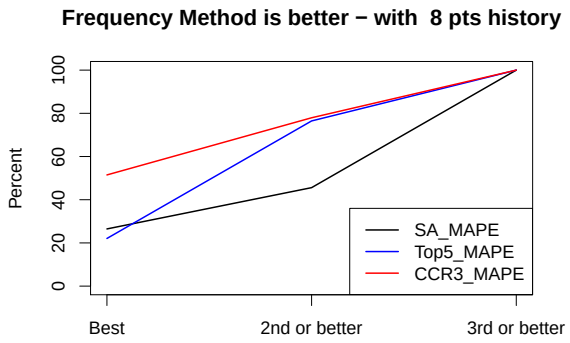
	Frequency % MAPE lowest of all methods								
	Economic forecasts					**Experimental data**			
Est. window	4 pts	8 pts	12 pts	16 pts	20 pts	4 pts	8 pts	12 pts	16 pts
SA	31.1	26.5	32.8	22.7	30.8	33.3	27.6	25.9	20.4
Top5	20.3	22.1	31.0	27.3	19.2	**35.0**	31.5	31.5	31.5
CCR3	**48.6**	**56.5**	**36.2**	**50.0**	**50.0**	31.7	**40.9**	**42.6**	**48.0**
Sample size	74	68	58	44	26	460	460	460	460

Table 15.: The frequency that the MAPE of a forecast combination applied to each of the data series is the lowest of all methods (⇑ better). Note: tied ranks are both scored as the max score which may result in columns that add to more than 100.

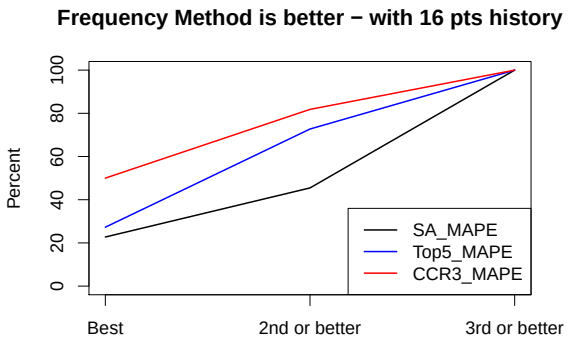
6.2.3 Risk performance

While it is important to understand which method has the lowest MAPE, a decision maker will want to understand the downside risk of using one particular method. Consider the frequency that a method places in second or better as plotted in Figure 30. In either Figure 30 a or b the CCR3 cumulative frequency of placing curve is always above that of Top5 and SA. This means that the likelihood of CCR3 performing better than Top5 or SA is higher regardless of whether the criteria is to finish first or no worse than 2nd. This is called stochastic dominance and suggests that CCR3 will always be a less risky choice than Top5 or SA for data similar to

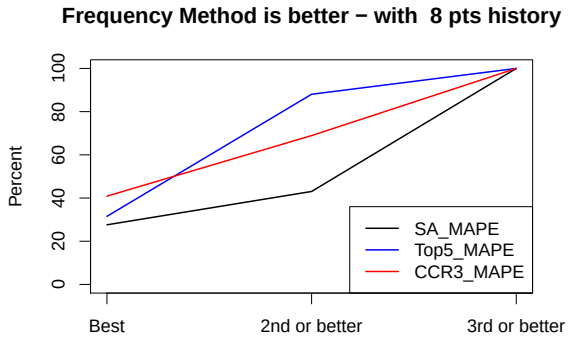
the economic data in this study. CCR3 is also statistically dominate over a simple average for the experimental data. But in this case Top5 does better if the criteria is finishing no worse than second place. We will look more closely at the differences between CCR3 and Top5 in Section 6.2.5.



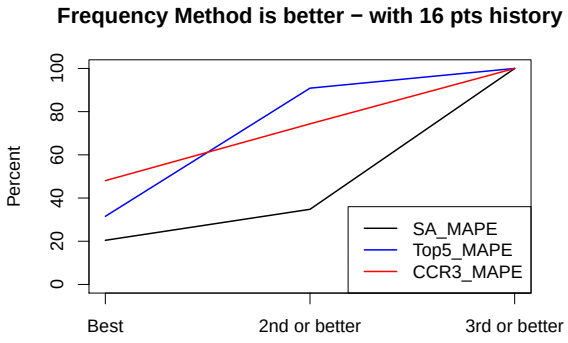
(a) Economic data with 8 est. points



(b) Economic data with 16 est. points



(c) Experimental data with 8 est. points



(d) Experimental data with 16 est. points

Fig. 30.: Cumulative frequency each method performs in the respective place or better (ie. 2nd or better is the frequency a method is in 1st or 2nd in relative performance to other methods).

6.2.4 Statistical performance comparison

The previous discussions have been based on average performance across entire sets of data from either a particular knowledge domain in the case of experimental data or from an economic indicator in a specific country over a period of time. We can use a sign test on the individual forecast points across knowledge domains and across economic indicators to determine if the median of the MAPE differences between the two methods evaluated at each point (a paired comparison) is statistically greater than zero (Dixon and Mood, 1946). The advantage of the sign test is that it is nonparametric and will be robust to the different distributions, levels, and outliers that we may have in the many data series in this study. The disadvantage is that it is not very sensitive and it does not provide any indication of how different the errors are. This lack of sensitivity will make it a conservative test. We will look at the economic and experimental forecasts separately as there appears to be a performance difference between the two data sets.

	CCR3 vs. SA MAPE Evaluated at each forecast point								
	Economic forecasts					**Experimental data**			
Est. window	4 pts	8 pts	12 pts	16 pts	20 pts	4 pts	8 pts	12 pts	16 pts
Mean MAPE diff.	1.67	5.47	4.28	8.22	12.33	14.80	17.20	18.50	19.30
Median MAPE diff.	0.21	0.27	0.12	0.22	0.20	1.80	2.70	3.15	3.37
Frequency better %	53.1	54.4	52.2	53.6	54.5	55.7	57.9	60.3	61.2
Pvalue sign test	**<0.00**	**<0.00**	**0.05**	**0.02**	**0.02**	**<0.00**	**<0.00**	**<0.00**	**<0.00**
N points	2261	1733	1300	877	541	9600	9600	9600	9600

Table 16.: Results of sign test comparing SA MAPE to CCR3 MAPE. A positive difference means that SA has a higher MAPE (⇑ better).

Table 16 shows the results of comparing SA MAPE to that of CCR3. In all

cases, the sign tests show that the median of the errors is not equal to zero at a high level of confidence. This implies that the alternative hypothesis of CCR3 is better, is likely true. To the author's knowledge, this is the first time an estimated weight has been shown to be statistically better than a simple average on a large body of data. (Note: Mannes et al. (2014) has shown that Top5, which is an inclusion approach, was statistically better than a simple average on similar data with 5 estimation points).

	CCR3 vs. Top5 MAPE Evaluated at each forecast point								
	Economic forecasts					**Experimental data**			
Est. window	4 pts	8 pts	12 pts	16 pts	20 pts	4 pts	8 pts	12 pts	16 pts
Mean MAPE diff.	(0.99)	1.37	0.43	3.76	2.10	(1.34)	(0.42)	(0.68)	(0.92)
Median MAPE diff.	0.10	0.07	(0.03)	0.08	(0.00)	(0.46)	(0.21)	(0.01)	(0.16)
Frequency better %	51.7	52.4	48.1	53.0	49.2	47.8	49.0	49.8	48.9
Pvalue sign test	**0.02**	**0.02**	0.87	**0.02**	0.53	1.00	0.98	0.58	0.98
N points	2261	1733	1300	877	541	9600	9600	9600	9600

Table 17.: Results of sign test comparing Top5 MAPE to CCR3 MAPE. A positive difference means that Top5 has a higher MAPE (⇑ better).

Table 17 shows the results of directly comparing Top5 MAPE to that of CCR3. In most cases, the economic data shows CCR3 is significantly better than Top5. The economic data is more positively correlated which will favor the CCR3 approach. However, on the experimental data set, it appears that Top5 performs better. Therefore a decision maker should use the CCR3 approach when combining economic data, but the Top5 approach may be better when combining expert opinions where there may be less correlation.

Table 18 shows the results of directly comparing SA MAPE to that of CCR3.B98. In this scenario, the CCR3 weights are only used if there is sufficient diversity in expert

	CCR3.B98 vs. SA MAPE evaluated at each forecast point								
	Economic forecasts					**Experimental data**			
Est. window	4pts	8pts	12pts	16pts	20pts	4pts	8pts	12pts	16pts
Mean MAPE diff.	0.28	1.66	1.33	4.30	11.67	14.80	17.20	18.50	19.30
Median MAPE diff.	0.00	0.07	0.13	0.16	0.19	0.00	0.42	1.43	1.80
Frequency better %	48.6	51.3	52.7	53.0	55.6	39.7	51.1	54.4	55.6
Pvalue sign test	0.38	0.09	**0.02**	**0.03**	**<0.00**	**<0.00**	**<0.00**	**<0.00**	**<0.00**
N points	2261	1733	1300	877	541	9600	9600	9600	9600

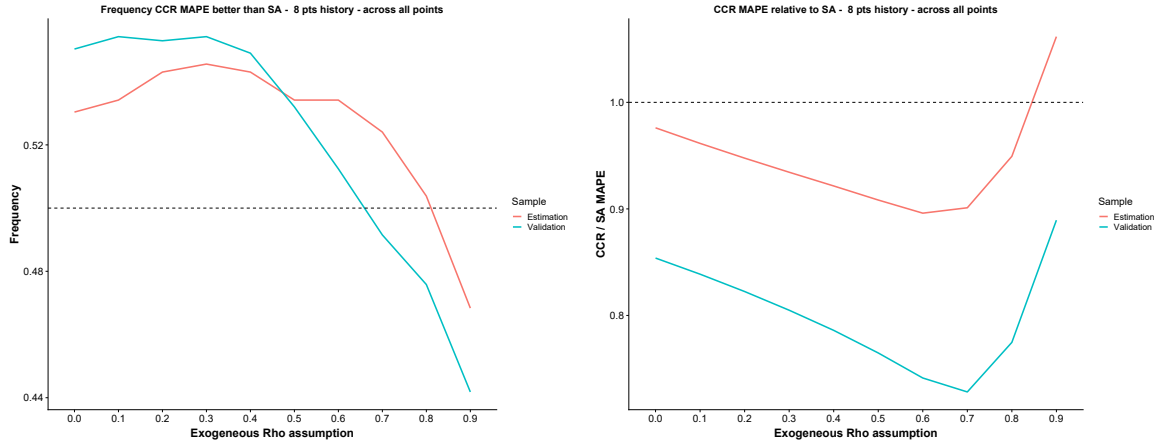
Table 18.: Results of sign test comparing SA MAPE to CCR3.B98 MAPE. A positive difference means that SA has a higher median MAPE ($\Uparrow$ better). Note in some cases the CCR3.B98 will chose a simple average, so that the frequency better need not be above 50% for the method to be better.

skill to justify estimating a weight. Otherwise a simple average is used. Similar to the results of SA versus CCR3, the CCR3.B98 is significantly better than SA. This shows that the critical skill ratio approach is not detrimental to the overall performance and may provide downside protection by avoiding some mis-estimation errors when a simple average would be better.

6.2.4.1 Robustness of an exogenous assumption for the level of correlation in CCR

The robustness of the exogenous assumption is illustrated in Figure 31. We split the economic data into two periods, the first half of the data points, chronologically, in a series were assigned to the estimation sample and the second half ere assigned to the validation sample. The the forecast combination performance of common correlation weights with exogenous assumptions from 0.0 to 0.9 were calculated for each data set independently. The shapes and optimum points of the frequency better than a

simple average curve and of the CCR3 MAPE as a ratio to that for a simple average are essentially the same. This shows that the optimal exogenous assumption for the common correlation is the same in each independent sample and therefore is robust to changes in the two samples.



(a) Frequency CCR3 better than a simple average (⇑ better)

(b) Ratio of CCR3 MAPE to a simple average (⇓ better)

Fig. 31.: CCR3 performance in two splits of the economic data evaluated with 8 points of history.

As a second test for robustness, an estimation sample of 30 data points was used to determine the level of expert correlation that would maximize the frequency that common correlation weights were better than a simple average (Table 19). This estimated correlation level was then used to assess the out of sample performance on the remaining points in the series. Only series with a total of more than 40 combined forecast points were used. The in sample performance of all seven series matched or exceeded that of a common correlation level exogenous assumption of 0.3; five of the seven series continued to match or exceed the out of sample performance of a 0.3 exogenous assumption. Also the overall performance across all series was stable in the estimation sample and in the validation sample. This suggests that individual

estimates of an optimal correlation level by series can be robust with 30 estimation points.

Country Indicator	**Estimation sample (N=30)**			**Validation sample**		
	Rho Est.	Frequency better than SA		Frequency better than SA		No. obs.
		Rho=0.3	Rho=Best	Rho=Best	Rho=0.3	
Argentina Economic Growth %	0.8	0.47	0.57	0.33	0.42	12
Argentina Exchange Rate	0	0.5	0.5	**0.43**	0.43	14
Brazil Economic Growth %	0.7	0.57	0.63	0.62	0.66	29
Brazil Exchange Rate	0.4	0.67	0.67	**0.53**	0.53	32
Brazil Inflation %	0.1	0.67	0.67	**0.81**	0.74	27
Mexico Economic Growth %	0	0.4	0.53	**0.41**	0.38	29
Mexico Exchange Rate	0.1	0.43	0.5	**0.54**	0.54	35
Average	0.3	0.53	0.58	**0.53**	0.53	

Table 19.: Out of sample performance of an estimated common correlation level that maximizes frequency better than SA on a 30 point estimation data set.

6.2.5 How different are CCR3 and Top5?

The similarity of the CCR3 and Top5 curves in Figure 30 and the inconsistent head-to-head performance in Table 17 raises the question: How different are the results of these two methods? Figure 32 illustrates a typical trend line for an SA, Top5, and CCR3 combined forecasts. Although Top5 and CCR3 use two entirely different approaches to arrive at a combined forecast, their trend lines follow each other much more closely than the trend line of a simple average. This suggests that they are using similar combinations of weights. Top5 puts all of the weight equally on the top 5 experts. For comparison purposes, Figure 33 plots the weight that CCR3 applies cumulatively to the five most skilled experts in the combination. It is surprising to see that CCR3 assigns a cumulative weight of ~ 1 to the 5 most skilled experts regardless of the number of experts when fewer than 20 experts are in the combination. Even with 76 experts in the combination, CCR3 places over 70% of the weight on the most accurate five experts. However the converse is not true: CCR3 does not assign high negative weights to the worse expert (Figure 34). As we saw in Figure 13, covariance optimal weights do not assign weights symmetrically to higher and lower skilled experts. An expert with a skill ratio of 2 will receive a proportionally higher weight than the lower weight an expert with a skill ratio of 0.5 receives. In addition, the distribution of expert skill ratios may be asymmetric with greater relative differences on the higher skill side, but smaller relative differences on the lower skilled side.

Top5 and CCR3 are the only two combination methods, known to the author, that have shown a statistically significant improvement over a simple average across a large body of data. Even thought they arrive at this result by different means, the end solution is to put essentially all of the weight on the top five experts. Top5 is

basically accounting for expert error correlation in an approximate way by placing all of the weight on the top five experts. The theory for CCR3's performance developed in this dissertation then also explains why Top5 performs so well.

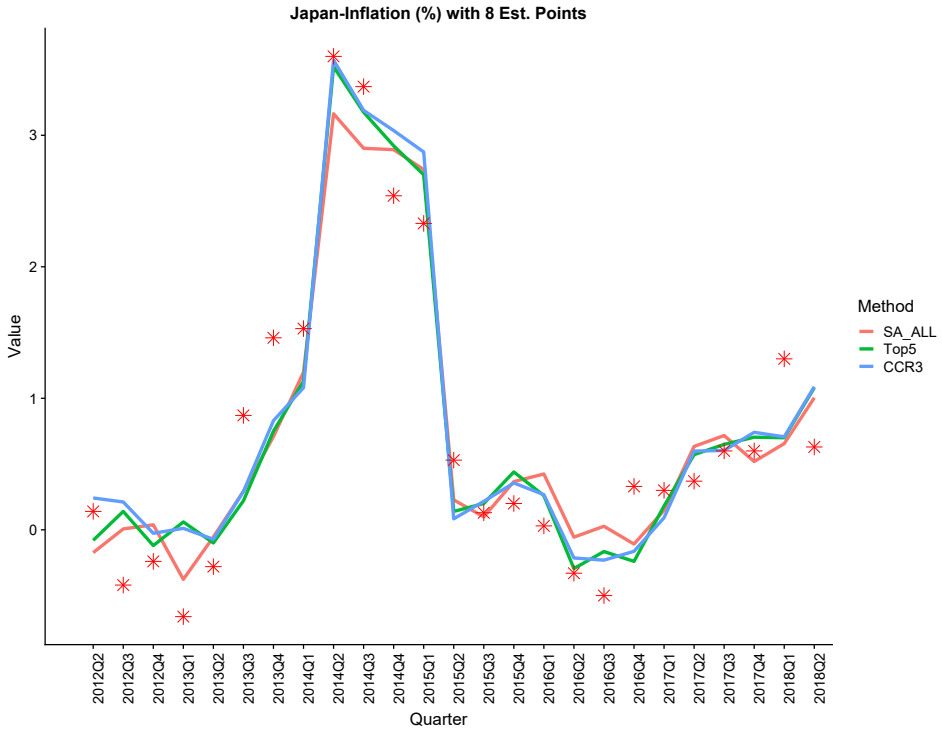
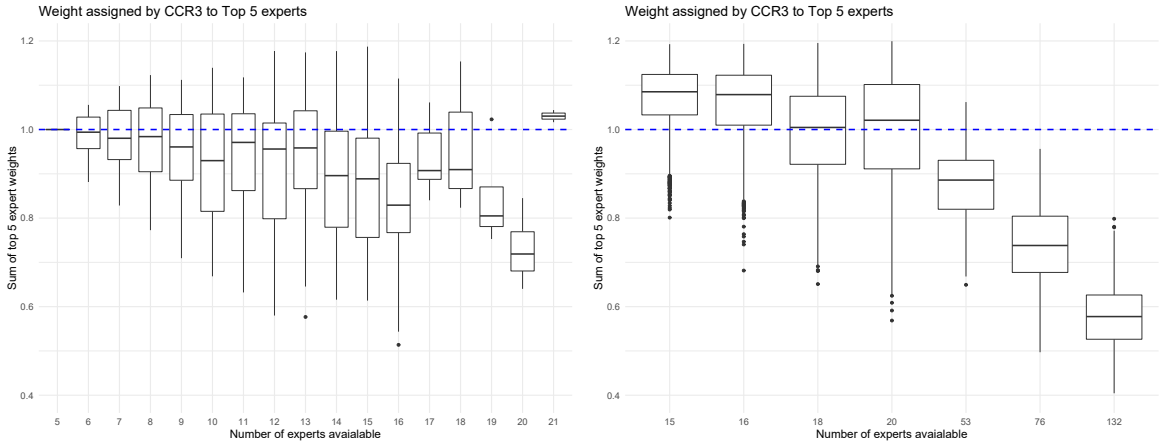
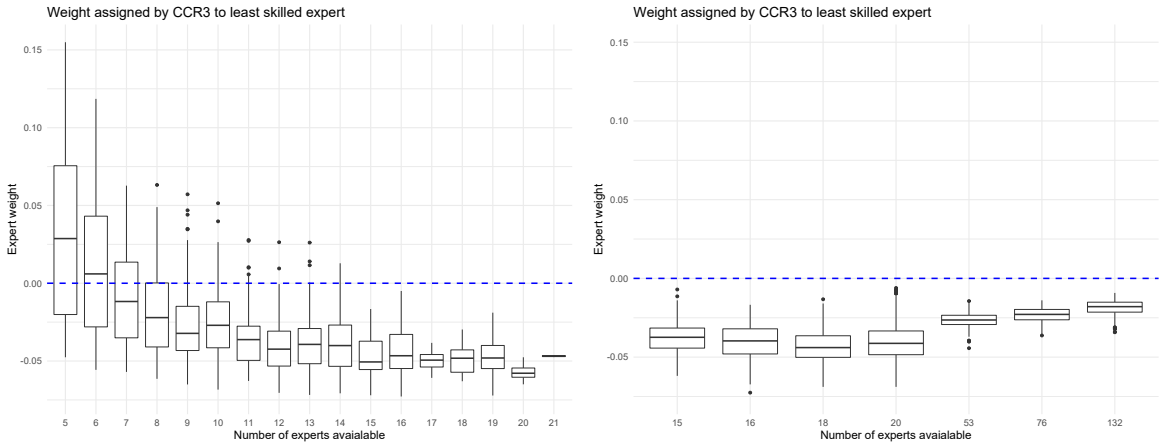


Fig. 32.: Forecast trends for Japanese inflation compared to realizations(red stars) .



(a) Economic data with 8 est. points (b) Experimental data with 8 est. points

Fig. 33.: Combined weight assigned to top5 experts by CCR3.



(a) Economic data with 8 est. points (b) Experimental data with 8 est. points

Fig. 34.: Minimum weight assigned by CCR3 method to experts not in top 5.

6.3 Discussion

The simulation study in Section 5.4 concluded that CCR3.B98 would be a better and more robust approach for combining forecasts. We have applied CCR3 and CCR3.B98 to two large and diverse bodies of data to test this conclusion and have found the following:

- **CCR3 performance** CCR3 performed statistically better than a simple average in all scenarios tested. In addition, compared to three other methods, it had the lowest MAPE in 8 out of 9 scenarios. This demonstrates that assuming an exogenous, low level of expert correlation improves estimated weights sufficiently to overcome estimation errors in a broad set of situations.

- **CCR3.B98** CCR3.B98 performed well only if there are 16 or more data points available to assess expert skill levels. This result validates the critical skill ratio threshold approach proposed in this dissertation. However, the practicality of this approach will be limited by the amount of historical data available.

- **Top5** The Top5 approach effectively accounts for expert error correlation and approximates the weights a CCR3 approach would use.

CHAPTER 7

CONCLUSIONS

We have taken a different approach to the Forecast Combination Puzzle than previous investigations and found that it is possible to beat a simple average. We developed a model which enabled analysis of the interactions of relative expert skill levels, number of experts, and expert correlations and their impacts on the estimation of weights. We investigated several new approaches (Table 20) to forecast combination that focused on minimizing weight estimation error. These proposals were evaluated in simulation studies and in empirical studies. The two approaches that beat a simple average are in bold. This analysis resulted in two new proposals for improved forecast aggregation and two new insights on multi-expert forecast combinations.

	All experts	**Select experts**
Fixed weights	- Simple average (Clemen, 1989)	- Top 5 experts (Mannes et al., 2014) - Drop experts below a critical threshold
Skill-based weights	-Variance based weights (Bunn, 1985) - Covariance based weights (Bates and Granger, 1969) **- Common correlation based weights**	- Drop experts below a critical threshold
Select method, fixed or skill-based weights	- Select method based on Akaike information criterion (Schmittlein et al., 1990) **- Select method using a critical threshold**	- Select method for each expert based on a critical threshold

Table 20.: Forecast combination methods evaluated in simulation study. Bold faced items were found to be significantly better than a simple average.

7.1 Summary of results

7.1.1 Critical skill thresholds

We defined and proposed the use of critical skill thresholds to determine when estimated weights are likely to perform better than the fixed weights of a simple average (Chapter 3). The purpose of this approach was to reduce estimation errors by only choosing to estimate weights when there is a high confidence that an estimated weight will be more accurate than a simple average. We investigated three ways to use critical skill ratio in the weight estimation process.

7.1.1.1 Choosing fixed versus skill based weights with critical skill ratios

The approach we called Best, only estimates weights for all experts when the skill ratio of one or more of the experts is beyond the critical skill ratio (high or low) at a prescribed level of confidence. In the simulation study with covariance, there was no discernible difference in accuracy performance of using the Best approach at critical skill ratios evaluated at 90 or 98% levels of confidence (Table 10, Best90 Best98). However, the Best approach at a critical skill ratio evaluated at a 98% confidence level reduced the frequency that combined forecast estimates were less than SA by almost 50% as compared to always using a common correlation weight (Table 11). On the economic data using the Best approach at a 98% confidence level performed noticeably worse than always using the common correlation approach when the estimation sample size was 12 or fewer points. However, the Best approach significantly improved performance with 20 estimation points. The Best approach at a 98% confidence level did a better job on the experimental data with no discernible impact as compared to always using a common correlation weight (CCR3). The Best approach at a 98% confidence level performance improved noticeable with 12

and greater estimation points. The difference in performance between the two data sets may be due to the more frequent low and negative expert correlations in the experimental data that would tend to favor the use of a simple average. These results demonstrate that using critical skill ratios can provide an improvement in forecast combination performance; however this approach is sensitive to sample size and perhaps other variables.

7.1.1.2 Selectively choosing which expert to assign a skilled based weight

A second application of critical skill ratios is to only assign a skill-based weight to the individual experts who have sufficiently differentiated skill as determined by the appropriate critical skill ratio. We called this the Select approach. In this case, the other undifferentiated experts would be assigned an average of the remaining weight. In this case, we minimize estimation error when there is little benefit from estimating weights. In the simulation study there was no discernible difference in this method over the simpler Best approach (Table 10 and 11). This is not entirely surprising as the skill-based weight for an undifferentiated expert is likely to be close to the average weight. Also, with many experts, the weight estimation error for similarly skilled experts will tend to average out. Based on these observations we continued our exploration using only the Best and common correlation approaches.

7.1.1.3 Dropping experts below a critical skill ratio threshold

A third application of critical skill ratios, is to inform the decision of which experts to keep in a forecast combination and which to exclude. The Top5 approach (Mannes et al., 2014) empirically found that keeping the five most skillful experts had the lowest forecast combination error. In general this is true, but are there not cases where the sixth expert is nearly as skilled as the fifth? We explored dropping

experts whose performance was below average as determined by the critical skill ratio. In the simulation study, dropping experts below a critical skill ratio evaluated at a 90% level of confidence performed better than Top5 when there were fewer than 10 experts in the combination, but Top5 performed better when there were 28 experts in the combination (Table 10). In the case of 28 experts the drop experts below the critical skill ratio method frequently retained 10 or more experts (Figure 25) versus the 5 experts used by Top5. All of the methods investigated that dropped experts performed consistently at a lower level of accuracy than either the common correlation approach or the Best98 approach. Based on these results, we did not further pursue investigations of dropping experts but kept the Top5 approach as a benchmark for comparison purposes.

Overall, the critical skill ratio approach can be used to reduce the variability in forecast combinations by reducing the frequency that a forecast combination performs worse than a simple average. This should give decision makers greater confidence in using skill-based estimates. Also, it only involves using a pre-estimated table to aid the decision process which should make it relatively easy to apply. However, the critical skill ratio approach is challenged by needing a larger sample size to be fully effective. Critical skill ratios are also sensitive to the levels of expert correlation (Figure 22). In our work we have not fully adjusted the critical skill ratio estimates for when the observed level of expert correlation is different than the level assumed by the weight estimation method.

7.1.2 Two versus multi-expert combinations

The three common weight estimation methods (SA, Var, Cov) all constrain the weights to add to 1. Then, in a two-expert combination there is only one degree of freedom, while in a mulit-expert combination there can be many degrees of freedom.

We have found that the behavior of estimated weights changes as a combination moves from one to many degrees of freedom. In a two-expert combination, increasing skill ratios decrease the standard deviation of an estimated weight, while in a multi-expert combination, the standard deviation of an estimated weight increases with increasing skill ratio. The intuition for this is on Figure 9. In Equation 5.1 we find that the weight of expert 1 is impacted by a $\rho(k-2)Skr_1$ term which is the level of expert correlation, times the number of experts minus 2, times the skill ratio of the expert. This term has the effect of increasing the weight's sensitivity to expert correlation when there are more than two experts in the combination. The interaction of the number of experts and degree of expert correlation can also be seen in Figure 15a where the threshold for when an expert's weight is positive changes from being convex to concave as the number of experts increases. Based on these observations, insights from two-expert combinations might not be relevant to extensions to multi-expert combinations. In case of multi-expert combinations, the multi-expert skill ratio defined and developed in this dissertation provides a new analytical approach to analyzing multi-expert combinations that includes the impacts of the greater degrees of freedom.

7.1.3 Covariance weights do work

Fifty years ago, Bates and Granger (1969) proposed using the covariance of expert errors in the estimation of optimal forecast combination weights. However, a simple average has been the dominate forecast combination approach as it avoids parameter estimation errors. We have proposed using covariance based weights with an exogenous assumption for the level of expert correlation as a way to reduce parameter estimation errors.

We found that this common correlation covariance method most frequently had

the lowest MAPE as compared to other methods in the economic data series for all historical estimation sample sizes analyzed (Table 15). In the experimental data series, it was most frequently the lowest MAPE for historical estimation sample sizes of 8 points or greater (Table 15). The next best method was a simple average in the economic data series and Top5 in the experimental data. In a direct paired comparison, we found that the MAPE for the common correlation covariance method was lower than the MAPE for a simple average on both the economic data series and on the experimental data set on a statistically significant basis (Table 16). Further analysis shows that the selection of an exogenous level of 0.3 for expert correlation was robust as it was replicated in an estimation and validation sample (Figure 31). We found that trying to estimate the level of expert correlation that maximizes the frequency a common correlation weight has a lower error than a simple average requires 30 or more data points to be robust. The availability of this much reliable history (over 7 years worth for quarterly forecasts) may not be practical; hence the preference for an exogenous assumption.

The setting of an exogenous level for the common expert correlation greatly simplifies the use of covariance based weights in two ways. First, this method is not sensitive to panel members entering or leaving the forecast combination as it does not depend on estimating relationships between the panel members. Frequently in economic forecast panels, individual experts will submit forecasts for some time periods and then not for others. This flexibility is an important advantage of using an exogenous assumption over any methods that involve estimating covariances. The exogenous assumption of the expert correlations also greatly simplifies the computation of the weights. With this method there is no need to invert a covariance matrix as there is a closed form solution for the inverted inter-class correlation matrix used by this approach. The use of an exogenous correlation assumption will improve the

performance of a forecast combination without a large increase in computational demands or complexity.

In this study we only looked at two possible exogenous assumptions for the common correlation level 0.3 and 0.5. A decision maker may choose to search for a more optimal value based on his historical data set. We found that a common correlation level of 0.3 was only marginally better than 0.5 in both, percent improvement (Table 8) and the risk of being higher than a simple average (Table 9). In addition, we found that using the minimum estimated expert pair correlation marginally, but consistently, performed better than using the average expert pair correlations on the same two performance metrics (Table 8 and 9). This suggests that there may not be a large advantage to estimating a better common correlation level. It further suggests that erring with a lower estimate of common correlation level may be beneficial. We believe that this may be due to the non-linear impact that correlation levels has on estimated weights (Figure 15) and likely on weight estimation errors.

7.1.4 Why five-expert combinations work well

Mannes et al. (2014) have shown that an average of the top 5 experts often performs better than the average of all experts in a combination. In addition, several other empirical studies have found that the improvement of a simple average combination over individual forecasts diminishes after 5 models are included (Clemen and Winkler, 1985, 1999; Makridakis, 1982; Newbold and Granger, 1974). We have found that even at low levels of expert correlation (0.3), the covariance optimum weight for a more highly skilled expert (skill ratio of 2) is several times (4x) that of the average expert in the crowd (Figure 14). We observed that even with as many as 20 experts, the covariance optimal weighting approach gives the top five experts a combined weight of $0.80 - 1.1$ (Figure 33). By definition, the Top5 combination method

achieves a similar expert weighting by giving the top five experts a combined weight of 1.0 and no weight to the remaining crowd. This suggests that Top5 is a simple way to achieve, approximately the same higher weighting that would be found by more complex covariance optimum weights.

7.2 Future work

The definition of a multi-expert critical skill ratio opened the door to several analyses and insights, but is a highly constrained model as only one expert is an outlier versus the crowd and all experts have the same level of error correlations. The model could be extended to answer the following questions and still be tractable. This path of research may develop into a heuristic to weight clusters of experts that is more effective than was tried by Genre et al. (2013).

1. **Heterogeneous levels of expert pair correlation** What if expert 1 (the outlier) had a different level of correlation with the crowd as well as a differing skill level? How would heterogeneous levels of correlation impact weights for a constant set of skill levels?

2. **Differing number of expert "outliers"** What if there were n experts in the outlier group all with similar skill ratios and correlations and $k - n$ experts in the rest of the crowd? How do the weight dynamics change based on the relative size of the outlier group to the crowd?

As a second path would be to improve upon the two new forecast combination heuristics. We have purposely kept the recommended aggregation heuristics simple both to benefit decision makers and also to provide clear insight into what effect is driving the improved performance. There is ample opportunity to refine and fine tune each heuristic. Ideas to consider are:

1. **What is the optimal level of assumed expert correlation?** In this study we tested $rho \in (0, 0.3, 0.5)$ to find an optimum level of exogenous correlation. We did not optimize further as we did not want to overly fit the model to the data but instead, show its broad practicality. In practice, decision makers may benefit from optimizing their choice of exogenous expert correlation on their historical data set. This is an optimization that could be updated on some periodic basis.

2. **Would other correlation estimation techniques perform better?** It is interesting that the minimum expert pair error correlation estimate performed better than the average in common correlation weights, and almost matched the performance of using an exogenous assumption in the simulation. This suggests that other estimation truncation approaches may be effective. Or perhaps, a Bayesian estimation approach with a strong prior.

3. **What is the optimal critical skill ratio threshold?** We used a 98% confidence level in the assessment based on the simulation results. Again, we did not want to over-fit the data. Further work on choosing the optimal confidence level would likely enhance the impact of this approach.

4. **Would using a broader selection of estimation approaches improve the critical skill ratio based decision approach?** Using a critical skill ratio to decide when to estimate weights (CCR3B98) worked well when there were sufficient estimation data points. Perhaps using a threshold to chose variance based weights and a further threshold to use CCR weights would achieve the higher performance of CCR3.

Appendix A

ANALYSIS OF EXPECTED VALUE AND VARIANCE IN A 2 FORECAST COMBINATION

Claeskens et al. (2016) developed the analysis of the expected value and variance of a forecast combination when the weights are estimated values. This appendix provides a summary of their derivation for a two forecast combination.

A.1 Two forecast system with fixed weights

Expected value for a two forecast combination with fixed weights:

$$\begin{aligned} f_c &= w_1 f_1 + w_2 f_2 \\ f_c &= w_1(\theta + e_1) + w_2(\theta + e_2) \\ f_c &= \theta + w_1 e_1 + w_2 e_2 \\ \mathbf{E}[f_c] &= \theta \end{aligned} \tag{A.1}$$

The variance for a two forecast combination with fixed weights:

$$\begin{aligned} Var(f_c) &= Var(\theta + w_1 e_1 + w_2 e_2) \\ Var(f_c) &= w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1 w_2 \rho \sigma_1 \sigma_2 \\ \text{Where: } \rho &= cov(e_1, e_2) \end{aligned} \tag{A.2}$$

A.2 Two forecast system with estimated weights

Expected value for a two forecast combination with estimated weights:

$$\begin{aligned}
\text{Let : } w_j &= \hat{w}_j + \epsilon_j \\
f_c &= \theta + \hat{w}_1 e_1 + \hat{w}_2 e_2 + \epsilon_1 e_1 + \epsilon_2 e_2 \\
\mathbf{E}[f_c] &= \theta + \mathbf{E}[\hat{w}_1 e_1] + \mathbf{E}[\hat{w}_2 e_2] + \mathbf{E}[\epsilon_1 e_1] + \mathbf{E}[\epsilon_2 e_2] \\
\mathbf{E}[f_c] &= \theta + Cov(\epsilon_1, e_1) + Cov(\epsilon_2, e_2) \qquad \text{(A.3)}
\end{aligned}$$

The variance for a two forecast combination with estimated weights:

$$
\begin{aligned}
Var(f_c) &= Var(\theta + \hat{w_1}e_1 + \hat{w_2}e_2 + \epsilon_1 e_1 + \epsilon_2 e_2) \\
Var(f_c) &= Var(\hat{w_1}e_1) + Var(\hat{w_2}e_2) + Var(\epsilon_1 e_1) + Var(\epsilon_2 e_2) + \\
&\quad 2Cov(\hat{w_1}e_1, \hat{w_2}e_2) + 2Cov(\hat{w_1}e_1, \epsilon_1 e_1) + 2Cov(\hat{w_1}e_1, \epsilon_2 e_2) + \\
&\quad 2Cov(\hat{w_2}e_2, \epsilon_1 e_1) + 2Cov(\hat{w_2}e_2, \epsilon_2 e_2) + 2Cov(\epsilon_1 e_1, \epsilon_2 e_2) \\
Var(f_c) &= \hat{w}_1^2\sigma_1^2 + \hat{w}_2^2\sigma_2^2 + 2\hat{w}_1\hat{w}_2\rho\sigma_1\sigma_2 + \\
&\quad \sigma_1^2(\hat{w}_1^2 + \sigma_{w1}^2) + Cov(\epsilon_1^2, e_1^2) - Cov(\epsilon_1, e_1)^2 + \\
&\quad \sigma_2^2(\hat{w}_2^2 + \sigma_{w2}^2) + Cov(\epsilon_2^2, e_2^2) - Cov(\epsilon_2, e_2)^2 + \\
&\quad 2\hat{w}_1(Cov(e_1, \epsilon_1 e_1) + Cov(e_1, \epsilon_2 e_2)) + \\
&\quad 2\hat{w}_2(Cov(e_2, \epsilon_1 e_1) + Cov(e_2, \epsilon_2 e_2)) + \\
&\quad 2Cov(\epsilon_1 e_1, \epsilon_2 e_2) \\
Var(f_c) &= \hat{w}_1^2\sigma_1^2 + \hat{w}_2^2\sigma_2^2 + 2\hat{w}_1\hat{w}_2\rho\sigma_1\sigma_2 + \\
&\quad \sigma_1^2(\hat{w}_1^2 + \sigma_{w1}^2) + Cov(\sigma_{w1}^2, \sigma_1^2) - Cov(\epsilon_1, e_1)^2 + \\
&\quad \sigma_2^2(\hat{w}_2^2 + \sigma_{w2}^2) + Cov(\sigma_{w2}^2, \sigma_2^2) - Cov(\epsilon_2, e_2)^2 + \\
&\quad 2\hat{w}_1(Cov(e_1, \epsilon_1 e_1) + Cov(e_1, \epsilon_2 e_2)) + \\
&\quad 2\hat{w}_2(Cov(e_2, \epsilon_1 e_1) + Cov(e_2, \epsilon_2 e_2)) + \\
&\quad 2Cov(\epsilon_1 e_1, \epsilon_2 e_2)
\end{aligned}
\tag{A.4}
$$

Appendix B

DISTRIBUTION OF VARIANCE BASED WEIGHTS IN A 2 FORECAST COMBINATION

Dickinson (1973) developed a closed form solution for the distribution of weights in a 2 forecast combination with no expert correlation as follows:

Let:

$w_i =$ the true variance based weight assigned to an expert i's forecast

$\hat{w}_i =$ the estimated weight assigned to an expert i's forecast based a sample size of$n$

$\sigma_i =$ the true error variance for expert i's forecasts

$k =$ the number of experts in the combination

Then, assuming independence we can directly determine the probabilistic distribution of a weight by observing that any one weight is based on the ratio of estimated

variances as follows:

$$
\begin{aligned}
w_i &= \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}} \text{ from maximum likelihood estimate - eq 2.2} \\
\hat{w}_i &= \frac{\frac{n-1}{\sigma_i^2 \chi^2(n)}}{\sum_{j=1}^{k} \frac{n-1}{\sigma_j^2 \chi^2(n)}} \text{substituting Chi squared distributions} \\
\hat{w}_i &= \frac{\frac{1}{\sigma_i^2 \chi^2(n)}}{\frac{1}{\sigma_i^2 \chi^2(n)} + \sum_{j=2}^{k} \frac{1}{\sigma_j^2 \chi^2(n)}} \\
\hat{w}_i &= \frac{1}{1 + \sum_{j=2}^{k} \frac{\sigma_i^2 \chi^2(n)}{\sigma_j^2 \chi^2(n)}} \\
\hat{w}_i &= \frac{1}{1 + \sum_{j=2}^{k} Z_j} \\
\text{Where : } Z_j &= \frac{\sigma_i^2 \chi^2(n)}{\sigma_j^2 \chi^2(n)} \text{ which is a scaled F-distributed variable}
\end{aligned} \tag{B.1}
$$

B.1 Proposition 1 - Scaled F-distribution

$$
\begin{aligned}
\text{If: } Z &= \frac{\sigma_1^2 \chi^2(n)}{\sigma_2^2 \chi^2(n)} \\
\text{Then: } Z &\propto \frac{\sigma_2^2}{\sigma_1^2} * F_{dist}(\frac{\sigma_2^2}{\sigma_1^2} * Z, n, n)
\end{aligned}
$$

Proof:

$$
\begin{aligned}
\text{Let: } X &= \frac{\chi^2(n)}{\chi^2(n)} \text{ an F-distributed random variable} \\
\text{Let: } Z &= \frac{\sigma_1^2}{\sigma_2^2} * X \\
\text{Then: } X &= \frac{\sigma_2^2}{\sigma_1^2} * Z \\
dX &= \frac{\sigma_2^2}{\sigma_1^2} * dZ \\
Z &\propto \frac{\sigma_2^2}{\sigma_1^2} * F_{dist}(\frac{\sigma_2^2}{\sigma_1^2} * Z, n, n) \text{ by variable transformation}
\end{aligned} \tag{B.2}
$$

B.2 Proposition 2 - Distribution of the optimal weight in a 2 Forecast ensemble

Proof:

$$\text{Let: } W_1 = \frac{\frac{n-1}{\sigma_1^2\chi^2(n)}}{\frac{n-1}{\sigma_1^2\chi^2(n)} + \frac{n-1}{\sigma_2^2\chi^2(n)}}$$

$$W_1 = \frac{\frac{\sigma_1^2\chi^2(n)}{\sigma_2^2\chi^2(n)}}{\frac{\sigma_2^2\chi^2(n)}{\sigma_1^2\chi^2(n)} + 1}$$

$$\text{Let: } Z = \frac{\sigma_2^2\chi^2(n)}{\sigma_1^2\chi^2(n)}$$

$$\text{Then: } W_1 = \frac{Z}{1+Z}$$

$$Z = \frac{W_1}{(1-W_1)} \tag{B.3}$$

$$dZ = \frac{1}{(1-W_1)^2} dW \tag{B.4}$$

Then by substitution into Equation B.2 and by a further variable transformation we have:

$$W_1 \propto \frac{\sigma_1^2}{\sigma_2^2} \frac{1}{(1-W_1)^2} F_{dist}\big(\frac{\sigma_1^2}{\sigma_2^2}\frac{W_1}{(1-W_1)}, n, n\big) \tag{B.5}$$

This expression can be further simplified by substituting into the pdf for the F-distribution as follows:

$$
\begin{aligned}
W_1 &\propto \frac{1}{Beta(\frac{v_1}{2},\frac{v_2}{2})}\big(\frac{v_2}{v_1}\big)^{\frac{V_2}{2}}\frac{\big(\frac{\sigma_1^2}{\sigma_2^2}\frac{W_1}{(1-W_1)}\big)^{(\frac{v_2}{2}-1)}}{\big(1+\frac{v_2\sigma_1^2}{v_1\sigma_2^2}\frac{W_1}{(1-W_1)}\big)^{\frac{v_1+v_2}{2}}}\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{(1-W_1)^2} \\
&\propto \frac{1}{Beta(\frac{v_1}{2},\frac{v_2}{2})}\big(\frac{\sigma_1^2 v_2}{\sigma_2^2 v_1}\big)^{\frac{V_2}{2}}\frac{\big(\frac{W_1}{(1-W_1))}\big)^{(\frac{v_2}{2}-1)}}{\big(1+\frac{v_2\sigma_1^2}{v_1\sigma_2^2}\frac{W_1}{(1-W_1}\big)^{\frac{v_1+v_2}{2}}}\frac{1}{(1-W_1)^2} \\
&\propto \frac{1}{Beta(\frac{v_1}{2},\frac{v_2}{2})}\big(\frac{\sigma_1^2 v_2}{\sigma_2^2 v_1}\big)^{\frac{V_2}{2}}\big(\frac{v_1\sigma_2^2}{v_1\sigma_2^2(1-W_1)+v_2\sigma_1^2W_1}\big)^{\frac{v_1+v_2}{2}}(W_1)^{(\frac{v_1}{2}-1)}(1-W_1)^{(\frac{(v_2)}{2}-1)}
\end{aligned}
\tag{B.6}
$$

In a forecasting situation one would expect to have the same number of historical data points n from each forecast to estimate the weights, this Equation would then further simplify to:

$$
W_1 \propto \frac{1}{Beta(\frac{n}{2},\frac{n}{2})}\big(\frac{\sigma_1^2}{\sigma_2^2}\big)^{\frac{n}{2}}\big(\frac{\sigma_2^2}{\sigma_2^2(1-W_1)+\sigma_1^2W_1}\big)^{(n)}(W_1)^{(\frac{n}{2}-1)}(1-W_1)^{(\frac{(n)}{2}-1)} \tag{B.7}
$$

In the case where the variances of each forecast are the same or nearly the same this distribution further simplifies to a Beta distribution with $(n-1)/2$ degrees of freedom:

$$
W_1 \propto \frac{1}{Beta(\frac{n}{2},\frac{n}{2})}(W_1)^{(\frac{n}{2}-1)}(1-W_1)^{(\frac{n}{2}-1)} \tag{B.8}
$$

B.3 R function to calculate weight density

The function makes a transformation of the F-distribution.

W = the weight of the first forecast at which the point density is being evaluated; the distribution is only defined on $W \in 0-1$

N = the number of points in the sample used to estimate the forecast variances

Ss1, Ss2 = the estimated variances of the two forecasts based on the sample

$$
\begin{aligned}
Wdensity &< -function(W, N, Ss1, Ss2)\{ \\
Z &< -(W/(1-W) * Ss1/Ss2) \\
X &< -df(Z, N, N) * (1-W)^{-}2 * Ss1/Ss2 \\
&return(X)\}
\end{aligned}
$$

Appendix C

IMPACT OF MULTIPLE EXPERTS, NO CORRELATIONS

First consider a combination with one outstanding expert (higher or lower skilled) and a number of average experts. Then from Equation 2.2:

$$w_i = \frac{\sigma_i^{-2}}{\sum_{j=1}^{k} \sigma_j^{-2}}$$

$$w_1 = \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sum_{j=2}^{k} \sigma_j^{-2}}$$

$$\sigma_{crowd}^{-2} = \frac{\sum_{j=2}^{k} \sigma_j^{-2}}{k-1} = \frac{1}{HA(\sigma_{2\ldots k})^2)}$$

Where HA() is the harmonic average.

$$w_1 = \frac{\sigma_1^{-2}}{\sigma_1^{-2} + (k-1)\sigma_{crowd}^{-2}}$$

$$\text{let: } Skr_1 = \frac{\sigma_1^{-2}}{\sigma_{crowd}^{-2}}$$

$$w_1 = \frac{Skr_1}{(Skr_1 - 1 + k)} \tag{C.1}$$

$$w_{crowd} = \frac{1}{(Skr_1 - 1 + k)} \tag{C.2}$$

$$w_1 - w_{fixed} = \frac{Skr_1}{(Skr_1 - 1 + k)} - \frac{1}{k}$$

$$w_1 - w_{fixed} = \frac{(Skr_1 - 1)(k-1)}{k^2 + k(Skr_1 - 1)} \tag{C.3}$$

Now consider the overall variance of the above combination:

$$
\begin{aligned}
\sigma^2_{combo} &= w_1^2\sigma_1^2 + (k-1)w^2_{crowd}\sigma^2_{crowd} \\
\sigma^2_{combo} &= w_1^2\frac{\sigma^2_{crowd}}{Skr_1} + (k-1)w^2_{crowd}\sigma^2_{crowd} \\
\sigma^2_{combo} &= (\frac{Skr_1}{Skr_1 - 1 + k})^2\frac{\sigma^2_{crowd}}{Skr_1} + \frac{(k-1)\sigma^2_{crowd}}{(Skr_1 - 1 + k)^2} \\
\sigma^2_{combo} &= \frac{Skr_1\sigma^2_{crowd}}{(Skr_1 - 1 + k)^2} + \frac{(k-1)\sigma^2_{crowd}}{(Skr_1 - 1 + k)^2} \\
\sigma^2_{combo} &= \frac{(Skr_1 - 1 + k)\sigma^2_{crowd}}{(Skr_1 - 1 + k)^2} \\
\sigma^2_{combo} &= \frac{\sigma^2_{crowd}}{Skr_1 - 1 + k} \qquad \text{(C.4)}
\end{aligned}
$$

We can also calculate the difference in variance of a weighted combination and a combination with equal weights:

$$
\begin{aligned}
\sigma^2_{delta} &= \sigma^2_{equalweights} - \sigma^2_{skillweights} \\
\sigma^2_{delta} &= \frac{\sigma^2_{crowd}}{k} - \frac{\sigma^2_{crowd}}{Skr_1 - 1 + k} \\
\sigma^2_{delta} &= \frac{\sigma^2_{crowd}(Skr_1 - 1)}{k^2 + k(Skr_1 - 1)} \\
\frac{\partial\sigma^2_{delta}}{\partial Skr_1} &= \frac{\sigma^2_{crowd}}{(Skr_1 + k - 1)^2} \qquad \text{(C.5)} \\
\frac{\partial\sigma^2_{delta}}{\partial k} &= -\frac{\sigma^2_{crowd}(Skr - 1)(2k + Skr - 1)}{k^2(k + Skr - 1)^2} \qquad \text{(C.6)}
\end{aligned}
$$

Appendix D

OPTIMAL WEIGHTS WITH ONE OVERALL COVARIANCE TERM

Consider a combination of forecasts from k experts. We will assume that expert 1 has a unique skill level and corresponding variance σ_1^2 while the $2 \dots k$ experts have relatively undifferentiated skill and a corresponding variance σ_{crowd}^2 as previously defined in Section XX. We will further assume that each expert on average is correlated with all of the other experts with an average correlation of ρ. Then the expert correlation matrix $\mathbf{A}$ is an inter-class matrix that can be inverted using the binomial inverse theorem with a known inverse (Press, 1972):

$$\mathbf{A} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1+(k-2)\rho}{C} & \frac{-\rho}{C} & \cdots & \frac{-\rho}{C} \\ \frac{-\rho}{C} & \frac{1+(k-2)\rho}{C} & \cdots & \frac{-\rho}{C} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-\rho}{C} & \cdots & \frac{-\rho}{C} & \frac{1+(k-2)\rho}{C} \end{bmatrix} \tag{D.1}$$

$$\text{where } C = 1 + (k-2)\rho - (k-1)\rho^2 \tag{D.2}$$

While the expert standard deviations can be placed in a diagonal matrix:

$$\mathbf{S} = \begin{bmatrix} \frac{\sigma_{crowd}}{\sqrt{Skr_1}} & 0 & 0 & \dots & 0 \\ 0 & \sigma_{crowd} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \sigma_{crowd} \end{bmatrix}$$

Finally the overall covariance matrix of the expert errors becomes:

$$\mathbf{\Sigma} = \mathbf{SAS}$$

$$\mathbf{\Sigma}^{-1} = \mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S}^{-1}$$

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{Skr_1+(k-2)\rho Skr_1}{C} & \frac{-\rho\sqrt{Skr_1}}{C} & \dots & \dots & \frac{-\rho\sqrt{Skr_1}}{C} \\ \frac{-\rho\sqrt{Skr_1}}{C} & \frac{1+(k-2)\rho}{C} & \frac{-\rho}{C} & \dots & \frac{-\rho}{C} \\ \vdots & \frac{-\rho}{C} & \ddots & \vdots & \vdots \\ \frac{-\rho\sqrt{Skr_1}}{C} & \vdots & \dots & \ddots & \frac{-\rho}{C} \\ \frac{-\rho\sqrt{Skr_1}}{C} & \frac{-\rho}{C} & \dots & \frac{-\rho}{C} & \frac{1+(k-2)\rho}{C} \end{bmatrix}$$

where $C = (1 + (k-2)\rho - (k-1)\rho^2)\sigma^2_{crowd}$

This inversion is valid as long as $C \neq 0$ which constrains $\rho \neq 1$. As ρ approaches 1 the inversion formula can become unstable.

We can then use the optimal weight formula in the presence of covariance as described in Section 2.1.

$$\vec{w}_{optimal} = \frac{\vec{1}^t\Sigma^{-1}}{\vec{1}^t\Sigma^{-1}\vec{1}}$$

where: $\vec{1}$ is a vector of k ones

$$Rowsum(\mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S}^{-1})_1^{-1} = \frac{Skr_1 + (k-2)\rho Skr_1 - (k-1)\rho\sqrt{Skr}_1}{C}$$

$$Rowsum(\mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S}^{-1})_{2\ldots k}^{-1} = \frac{1-\rho\sqrt{Skr_1}}{C}$$

$$\vec{1}^t(\mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S}^{-1})\vec{1} = \frac{Skr_1 + (k-1) + \rho(k-2)Skr_1 - 2\rho(k-1)\sqrt{Skr_1}}{C}$$

$$w_1 = \frac{Skr_1 + (k-2)\rho Skr_1 - (k-1)\rho\sqrt{Skr}_1}{Skr_1 + (k-1) + \rho(k-2)Skr_1 - 2\rho(k-1)\sqrt{Skr_1}}$$

$$w_1 = \frac{Skr_1 + \rho[(k-2)Skr_1 - (k-1)\sqrt{Skr}_1]}{Skr_1 + (k-1) + \rho[(k-2)Skr_1 - 2(k-1)\sqrt{Skr_1}]}$$

$$w_{2\ldots k} = \frac{1-\rho\sqrt{Skr_1}}{Skr_1 + (k-1) + \rho[(k-2)Skr_1 - 2(k-1)\sqrt{Skr_1}]}$$

REFERENCES

Aksu, C and S Gunter (1992). "An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts". In: *International Journal of Forecasting* 8.1, pp. 27–43.

Armstrong, J Scott (2001). "Combining forecasts". In: *Principles of forecasting*. Springer, pp. 417–439.

Bates, J. M. and C. W. J. Granger (1969). "The Combination of Forecasts". In: *Operations Research Quarterly* 20.4, pp. 451–468.

Blanc, S. and T. Setzer (2016). "When to choose the simple average in forecast combination". In: *Journal of Business Research* 69.10, pp. 3951–3962.

Bunn, Derek (1985). "Statistical efficiency in the linear combination of forecasts". In: *International Journal of Forecasting* 1.2, pp. 151–163.

Bunn, Derek W. (1981). "Two Methodologies for the Linear Combination of Forecasts". In: *The Journal of the Operational Research Society* 32.3, pp. 213–222.

Chan, Felix and Laurent L. Pauwels (2018). "Some theoretical results on forecast combinations". In: *International Journal of Forecasting* 34.1, pp. 64–74.

Claeskens, Gerda et al. (2016). "The forecast combination puzzle: A simple theoretical explanation". In: *International Journal of Forecasting* 32.3, pp. 754–762.

Clemen, Robert (1986). "Calibration and the aggregation of probabilities". In: *Management Science* 32.3, pp. 312–314.

— (1987). "Combining overlapping information". In: *Management Science* 33.3, pp. 373–380.

— (1989). "Combining forecasts: A review and annotated bibliography". In: *International journal of forecasting* 5.4, pp. 559–583.

Clemen, Robert and Robert Winkler (1985). "Limits for the precision and value of information from dependent sources". In: *Operations Research* 33.2, pp. 427–442.

— (1999). "Combining probability distributions from experts in risk analysis". In: *Risk analysis* 19.2, pp. 187–203.

Conflitti, Cristina, Christine De Mol, and Domenico Giannone (2015). "Optimal combination of survey forecasts". In: *International Journal of Forecasting* 31.4, pp. 1096–1103.

Croushore, Dean (1993). "Introducing: The Survey of Professional Forecasters". In: *Business Review*, pp. 3–15.

de Menezes, L. M. and Derek Bunn (1998). "The persistence of specification problems in the distribution of combined forecast errors". In: *International Journal of Forecasting* 14.3, pp. 415–426.

Dickinson, J. P. (1973). "Some Statistical Results in the Combination of Forecasts". In: *Operational Research Quarterly (1970-1977)* 24.2, pp. 253–260.

Diebold, Francis X (1989). "Forecast combination and encompassing: Reconciling two divergent literatures". In: *International Journal of Forecasting* 5.4, pp. 589–592.

Dixon, W. J. and A. M. Mood (1946). "The Statistical Sign Test". In: *Journal of the American Statistical Association* 41.236, pp. 557–566.

Elliott, Graham (Apr. 2011). "Averaging and the Optimal Combination of Forecasts". In:

Genre, Véronique et al. (2013). "Combining expert forecasts: Can anything beat the simple average?" In: *International Journal of Forecasting* 29.1, pp. 108–121.

Guerard J. B., J. and R. T. Clemen (1989). "Collinearity and the use of latent root regression for combining GNP forecasts". In: *Journal of Forecasting* 8.3, pp. 231–238.

Gunter, Sevket (1992). "Nonnegativity restricted least squares combinations". In: *International Journal of Forecasting* 8.1, pp. 45–59.

Hsiao, Cheng and Shui Ki Wan (2014). "Is there an optimal forecast combination?" In: *Journal of Econometrics* 178, pp. 294–309.

Hurvich, Clifford M. and Chih-Ling Tsai (1989). "Regression and Time Series Model Selection in Small Samples". In: *Biometrika* 76.2, pp. 297–307.

Jose, Victor Richmond R. and Robert L. Winkler (2008). "Simple robust averages of forecasts: Some empirical results". In: *International Journal of Forecasting* 24.1, pp. 163–169.

Kang, Heejoon (June 1986). "Unstable Weights in the Combination of Forecasts". In: *Management Science* 32.6, p. 683.

Kendall, Maurice G (1962). *Advanced Theory of Statistics*. New York : Hafner Publ. Co.

Larrick, Richard P., Katherine A. Burson, and Jack B. Soll (2007). "Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)". In: *Organizational Behavior and Human Decision Processes* 102.1, pp. 76–94.

Makridakis, S. (1982). "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition." In: *Journal of Forecasting* 1.2, pp. 111–153.

Mannes, Albert E., Jack B. Soll, and Richard P. Larrick (2014). "The wisdom of select crowds". In: *Journal of Personality and Social Psychology* 107.2, pp. 276–299.

Maor, Eli (1977). "A MATHEMATICIAN'S REPERTOIRE OF MEANS". In: *The Mathematics Teacher* 70.1, pp. 20–25.

Newbold, P. and C. W. J. Granger (1974). "Experience with Forecasting Univariate Time Series and the Combination of Forecasts". In: *Journal of the Royal Statistical Society. Series A (General)* 137.2, pp. 131–165.

Palm, Franz C and Arnold Zellner (1992). "To combine or not to combine? Issues of combining forecasts". In: *Journal of Forecasting* 11.8, pp. 687–701.

Press, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart, and Winston, p. 23.

Schmittlein, David C, Jinho Kim, and Donald G Morrison (1990). "Combining forecasts: Operational adjustments to theoretically optimal rules". In: *Management Science* 36.9, pp. 1044–1056.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Smith, Jeremy and Kenneth F. Wallis (2009). "A Simple Explanation of the Forecast Combination Puzzle*". In: *Oxford Bulletin of Economics and Statistics* 71.3, pp. 331–355.

Soll, Jack B. and Richard P. Larrick (2009). "Strategies for revising judgment: How (and how well) people use others' opinions". In: 35.3, pp. 780–805.

Soll, Jack B. and Albert E. Mannes (2011). "Judgmental aggregation strategies depend on whether the self is involved". In: *International Journal of Forecasting* 27.1, pp. 81–102.

Stock, James and Mark Watson (June 1998). *A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series*. Working Paper 6607. National Bureau of Economic Research.

— (2004). "Combination Forecasts of Output Growth in a Seven-Country Data Set". In: *Journal of Forecasting* 23, pp. 405–430.

Winkler, Robert (1981). "Combining Probability Distributions from Dependent Information Sources". In: *Management Science* 27.4, pp. 479–488.

Winkler, Robert and Robert Clemen (1992). "Sensitivity of weights in combining forecasts". In: *Operations Research* 40.3, pp. 609–614.

Winkler, Robert and Spyros Makridakis (1983). “The Combination of Forecasts”. In: *Journal of the Royal Statistical Society. Series A (General)* 146.2, pp. 150–157.

VITA

David Patterson Soule was born on Sep. 29, 1957 n Upper Montclair, NJ and is an American citizen. He graduated from Montclair Kimberley Academy in 1975, Montclair, NJ. He received his Bachelor of Science degree in Mechanical Engineering from the Massachusetts Institute of Technology, Cambridge, MA in 1979. He subsequently received a Masters in Business Administration with distinction from New York University, NY, NY in 1981. He has held senior operations, supply chain, and analytical roles in chemical manufacturing, transportation, and financial services. He retired from Capital One Financial in 2018. He is a certified Six Sigma Master Black Belt and a Bronze Lean practitioner. He is currently a visiting lecturer at the Robbins School of Business, University of Richmond, Richmond, VA. His conference presentations include:

- Forecast aggregation with spatial auto-correlation - INFORMS Advances in Decision Analysis, Austin, TX 2017

- Forecast aggregation with spatial auto-correlation- American Meteorological Society 24th Conference on Probability and Statistics, Baltimore MD, 2017