Forecast Combination for Outlier Protection and Forecast Combination Under Heavy Tailed Errors

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Gang Cheng

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Yuhong Yang, Adviser

November 2014

Acknowledgements

I am thankful to my advisor, Professor Yuhong Yang, for leading me into the research area of forecast combination, for guiding me to find specific interesting research topics, for inspiring me to explore problems in different angles, for encouraging me to propose new ideas to important problems, for training my scientific writing skills patiently, and for all the great discussions and advise from which I had benefited a lot in my study in the Ph. D program and from which I will benefit for sure in the rest of my life. Thank you, Professor Yang.

I am thankful to Professor Snigdhansu Chatterjee, Professor Adam Rothman from School of Statistics, University of Minnesota at Twin cities and Professor William Li from Carlson School of Management, University of Minnesota at Twin cities for serving in my defense committee, for taking time reviewing my thesis, and for providing constructive suggestions and comments to my research.

I am thankful to my school, School of Statistics, University of Minnesota at Twin cities, for providing such a great environment that supports my study and life in the past few years in many aspects. I really enjoyed the conversations and friendship with the faculty members, the staffs, and friends in the school.

Finally, I thank my parents. Without their encouragement and vision, I can not arrive at this moment. I thank my wife Qi Yan and daughter Ella, for their love and support, and for making my life more meaningful and joyful.

Dedication

This dissertation is dedicated to my parents: Zhian Cheng and Shaoying Xu.

Abstract

Forecast combination has been proven to be a very important technique to obtain accurate predictions. Numerous forecast combination schemes with distinct properties have been proposed. However, to our knowledge, little has been discussed in the literature on combining forecasts with minimizing the occurrence of forecast outliers in mind. An unnoticed phenomenon is that robust combining, which often improves predictive accuracy (under square or absolute error loss) when innovation errors have a tail heavier than a normal distribution, may have a higher frequency of prediction outliers. Given the importance of reducing outlier forecasts, it is desirable to seek new loss functions to achieve both the usual accuracy and outlier-protection simultaneously. In the second part of this dissertation, we propose a synthetic loss function and apply it on a general adaptive theoretical and numeric results support the advantages of the new method in terms of providing combined forecasts with relatively fewer large forecast errors and comparable overall performances.

For various reasons, in many applications, forecast errors exhibit heavy tail behaviors. Unfortunately, to our knowledge, little has been done to deal with forecast combination for such situations. The familiar forecast combination methods such as simple average, least squares regression, or those based on variance-covariance of the forecasts, may perform very poorly in such situations. In the third part of this dissertation, we propose two forecast combination methods to address the problem. One is specially proposed for the situations that the forecast errors are strongly believed to have heavy tails that can be modeled by a scaled Student's t-distribution; the other is designed for relatively more general situations when there is a lack of strong or consistent evidence on the tail behaviors of the forecast errors due to shortage of data

and/or evolving data generating process. Adaptive risk bounds of both methods are developed. Simulations and a real example show the excellent performance of the new methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Forecasting is widely and regularly used to help with decision making in many areas of our modern life. Because of the availability of different sources of information, different methods and distinct backgrounds or preferences of the forecasters, multiple forecasts are available for the target variable of interest in many applications. In order to get the most accurate forecasts by taking advantage of the candidate forecasts, the strategy of forecast combination is often applied. Forecast combination approaches generate new forecasts by combining all or some of the candidate forecasts.

1.1 Combination and Selection

When multiple forecasts are available for a target variable, there are two popular approaches to obtain forecasts with competitive predictive performance under predetermined evaluation criteria. One approach picks a best forecast under certain measures and the other creates a new forecast by combining all or some of the candidate forecasts in certain ways.

In order to pick a best forecast in the candidate pool, many model selection methods have been proposed in the past several years. Unfortunately, to our best knowledge, no selection method is found to outperform the others universally. So the

users have to pick proper selection methods among a large pool of options for their statistical modeling problems, which can be really challenging.

As an alternative to the approach of selecting a forecast from many candidates, the forecast combination approach takes advantage of all or some of the individual forecasts to create a new forecast. Literature shows that well designed forecast combination methods can often outperform the best individual forecaster, as demonstrated in applications such as tourism, wind power generation, finance and economics in the last fifty years.

1.2 Popular Forecast Combination Methods

Since the seminal work of forecast combination by Bates & Granger (1969), thousands of research papers have been published on this topic with various combining schemes. We will introduce the framework of some of the most popular methods. Before that, let us introduce some notations.

Suppose we have N candidate forecasters, and combination starts at time n_0 (the first $n_0 - 1$ observations are used as training data). Let $\hat{y}_{j,i}$ be the forecast of y_i from candidate forecaster j. Let the combined forecast for y_i from method Δ be $\hat{y}_i^{\Delta}$ and $w_{i,j}^{\Delta}$ be the combination weight of $\hat{y}_{j,i}$.

Let the true model be $y_i = m_i + \epsilon_i$, where $m_i = E(y_i)$ and ϵ_i follows a certain distribution. Let σ_i^2 be the conditional variance of ϵ_i, if its variance exists.

- Combining via simple averaging (e.g., Stock & Watson, 1999). This method simply puts: $w_{i,j} = \dfrac{1}{N}$ for all $i \geq n_0$ and $1 \leq j \leq N$. There are many variations of this method. For example, when some of the candidate forecasts have outliers, then taking the mean of $\{\hat{y}_{i,1}, \cdots, \hat{y}_{i,N}\}$ as the combined forecast for y_i is less robust than taking the median for any $i \geq n_0$. Another variation designed to handle the potential outliers is the trimmed mean method (e.g., Wei

& Yang, 2012) which removes some of the largest and/or the smallest ones from $\{\hat{y}_{i,1}, \cdots, \hat{y}_{i,N}\}$ before taking the simple average as the combined forecast for y_i. In this dissertation, the simple average, median and trimmed mean methods are denoted as `SA, MD` and `TM`, respectively.

- Combining via variance-covariance estimation of the candidate forecasts (e.g., Bates & Granger, 1969). Let the covariance matrix of the candidate forecasts be Σ and the candidate forecasts be $f := (f_1, \cdots, f_N)$ and are unbiased. Then the optimal combining weight vector $w := (w_1, \cdots, w_N)$ satisfies:

$$\begin{aligned} w &= \arg\min \mathrm{Var}(w^T f) = \arg\min w^T \Sigma w \\ &s.t.\ ||w||_1 = 1,\ w_j \geq 0\ \forall 1 \leq j \leq N, \end{aligned}$$

where $||w||_1 = \sum_{j=1}^{N} |w_j|$ is the L_1-norm of w. Since Σ is generally unknown and also not easy to estimate, the solution of the above optimization problem is approximated by:

$$w_{i,j} = \frac{\frac{1}{\hat{\sigma}^2_{i,j}}}{\sum_{j'=1}^{N} \frac{1}{\hat{\sigma}^2_{i,j'}}},$$

where $\hat{\sigma}^2_{i,j}$ is the estimated conditional variance of σ_i^2 for forecaster j. If there is no specific information of the models behind the candidate forecasts, there are three general ways to calculate $\hat{\sigma}^2_{i,j}$: One uses the sample variance of $\{y'_i - \hat{y}_{i',j}\}_{i'=1}^{i-1}$ which takes advantages of all the previous information, one sets $i - i'$ to be a constant (e.g., $i - i' = 20$) which uses a rolling window (with fixed length) of data to do the calculation, and the last one is a compromise between the first two methods: It can use all the previous information but gives the newer information more emphasis in calculating $\hat{\sigma}^2_{i,j}$. That is: $\hat{\sigma}^2_{i,j} = \frac{\sum_{i'=1}^{i} \rho^{i-i'}(y'_i - \hat{y}_{i',j})^2}{\sum_{i'=1}^{i} \rho^{i-i'}}$, where $0 < \rho \leq 1$. See, e.g., Stock & Watson (2003), for details.

In this dissertation, the first way to calculate $\hat{\sigma}_{i,j}^2$ in this method is denoted as `BG`.

- Combining via Bayesian model averaging (e.g., Min & Zellner, 1993). The idea is that each of the N candidate models has a certain probability to be the true model and the combined forecast is the expectation (weighted mean) of the candidate forecasts. When new information arrives, the probability updates in the Bayesian framework. A popular approximation of this method is:

$$w_{i,j} = \frac{\exp(-\widehat{BIC}_{i,j})}{\sum_{j'=1}^{N} \exp(-\widehat{BIC}_{i,j'})},$$

where $\widehat{BIC}_{i,j}$ is the estimated Bayesian information criterion (BIC) of forecaster j at time point $i-1$.

- Combining via regression on candidate forecasts (e.g., Granger & Ramanathan, 1984). That is,

$$w_{i,j} = \arg \min_{(\beta_0,\beta_1,\cdots,\beta_N)} \sum_{i'=1}^{i-1} (y'_i - \beta_0 - \sum_{j'=1}^{N} \beta_{j'} \hat{y}_{i',j'})^2, \quad \hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{N} w_{i,j} \hat{y}_{i,j}.$$

There is one constrained version of this method: $\beta_0 = 0$, $w_{i,j} \geq 0$ and $\sum_{j=1}^{N} w_{i,j} = 1$ for each $i \geq n_0$. The ordinary linear regression method and the constrained version are denoted as `LR` and `CLR`, respectively, in this dissertation. Because `LR` can be very unstable, an approach which is roughly a mix of `SA` and `LR` is proposed. That is: $w_{i,j} = (1-\lambda)\frac{1}{N} + \lambda w_{i,j}^{\texttt{LR}}$, where $w_{i,j}^{\texttt{LR}}$ is the weight from `LR` without intercept ($\beta_0 = 0$), where $\lambda = \max\{0, 1 - \kappa\frac{N}{i-2-N}\}$ with κ could be estimated via empirical Bayes methods. See, e.g., Stock & Watson (2004), for details.

- Combining via exponential re-weighting (e.g., Yang, 2004). The AFTER method proposed in (Yang, 2004) is:

$$w_{i,j} = \frac{\frac{1}{\hat{\sigma}_{i,j}}\phi\left(\frac{y_i - \hat{y}_{i,j}}{\hat{\sigma}_{i,j}}\right)}{\sum_{j'=1}^{N}\frac{1}{\hat{\sigma}_{i,j'}}\phi\left(\frac{y_i - \hat{y}_{i,j'}}{\hat{\sigma}_{i,j'}}\right)},$$

  where $\phi(\cdot)$ is the probability density function of a standard normal distribution and $\hat{\sigma}_{i,j}$ is an estimator of σ_i from forecaster j.

Reviews and discussions of the research results are available in Clemen (1989), Newbold & Harvey (2002), Timmermann (2006) and Lahiri et. al (2013).

1.3 Combining for Adaptation and Improvement

Forecast combination methods can be categorized into two groups by their goals of combination. In one group, the goal is to build a forecast that outperforms all the candidate models while the other one aims at adapting the performance of the best individual forecasts automatically (see, Yang, 2004).

The price that the methods in the first group, such as `LR` and `CLR`, have to pay is that their convergence rate is materially lower than the methods in the second group (such as AFTER method from Yang (2004) and Wei & Yang (2012)). See, e.g., Tsybakov (2003) and Yang (2004), for more detailed discussions.

So when the performances of the the best individual forecasts are acceptable, aiming at adapting the performance of the best candidate forecast is more feasible in practice Notice that the best individual forecasts can change over time, so automatically adapting the performance of the best individual forecast over time can make the adaptively combined forecasts eventually outperform all the individual forecasts. However, if all candidate forecasts are weak in that they provide predictive perfor-

mances which are far from acceptable, then improvement is more reasonable than adaptation.

1.4 Structure of This Dissertation

Although many forecast combination methods have been proposed in the fast few decades, to our knowledge, little has been discussed in the literature on combining forecasts with minimizing the occurrence of forecast outliers in mind. An unnoticed phenomenon is that robust combining, which often improves predictive accuracy (under square or absolute error loss) when innovation errors have a tail heavier than a normal distribution, may have a higher frequency of prediction outliers. Given the importance of reducing outlier forecasts, it is desirable to seek new loss functions to achieve both the usual accuracy and outlier-protection simultaneously. In chapter 2, we will discuss the roles of loss functions in forecast combination methods and propose AFTER based combination methods to generate forecasts that are outlier protective. That is, the combined forecasts are more likely to have fewer large forecast errors than all of the candidate forecasts. Theoretical and numerical results support the advantages of the proposed methods well.

For various reasons, in many applications, forecast errors exhibit heavy tail behaviors. Unfortunately, to our knowledge, little has been done to deal with forecast combination for such situations. In chapter 3, we propose two AFTER-based forecast combination approaches. One is designed for the situation that there is strong evidence that the errors are heavy-tailed, while the other is designed for situations when moderate evidence for heavy-tailed error distributions exists. The risk bounds of both methods are derived. Systematic numeric simulation and real data analysis shows the advantage of proposed methods consistently.

Chapter 4 describes an associated R package for the methods discussed in Chapters

1, 2 and 3. It not only shows programming details of the combination methods but also provides practical suggestions to the users to pick the proper combination methods and use them in the right way.

Chapter 5 concludes the dissertation and the proofs of the theorems are provided in the appendix.

Chapter 2

Forecast Combination with Outlier Protection

2.1 Introduction

Loss functions play important roles in forecast combination in two intertwining directions: they may serve as a key ingredient in combination formulas and they are used to define performance evaluation criteria. Take forecast combination via ordinary least squares regression for example, the combining weights of the forecasts are trained by minimizing the sum of the squared errors (the L_2-loss), while the performance of the combined forecasts can also be evaluated under the same loss function or a different one such as the L_1-loss.

Indeed, the use of a loss function in the first direction is found in many popular combination schemes, such as the regression based combination (e.g., Bates & Granger, 1969; Granger & Ramanathan, 1984) and many adaptive/recursive forecast combination schemes (e.g., Yang, 2000, 2004; Zou & Yang, 2004; Wei & Yang, 2012). Take the L_1-loss in the L_1-AFTER of Wei & Yang (2012) for example, it uses the cumulative L_1-loss to summarize the historical performance of the candidate forecasts to decide the combining weights for predicting the next observation.

The need to use loss functions in the second direction is obvious. The objec-

tive of any combination strategy is to provide forecasts to better serve some predefined/predetermined goals, which are often characterized in terms of loss or utility functions. While the symmetric quadratic loss is most often used in both the theoretical and empirical research works, other loss functions have been explored for forecast combination (see e.g., Zeng & Swanson, 1998; Elliott & Timmermann, 2004; Pai & Lin, 2005; Chen & Yang, 2007; Wei & Yang, 2012). In particular, in fields such as economics and finance, asymmetric evaluation criteria are important to study (see e.g., Zellner, 1986; Granger & Newbold, 1986; West et. al, 1997; Christoffersen & Diebold, 1997; Granger & Pesaran, 2000; Diebold, 2001). In our context, for example, the linex loss, lin-lin loss and asymmetric squared loss functions are discussed in detail as forecast performance evaluation criteria in Elliott & Timmermann (2004).

Besides the loss functions mentioned above, the frequency of large forecast errors (larger than some thresholds in the positive or negative directions) is also important since decisions made for the future based on substantially over or under forecasting may cause severe undesirable consequences. For instance, a severe forecast error on demand may lead to a company's drastic over or under production, negatively affecting its profit. In spite of the obvious importance of having minimal frequency of large forecast errors, to our knowledge, little has been discussed in the literature on combining strategies with a control on the occurrence of large forecast errors directly. It is clear that optimization under the L_2-, L_1-loss or other performance measures can have some effect on the control of the frequency of large forecast errors, but the control is not explicit. It is thus of interest to understand how the different loss functions perform in forecast combination with respect to the occurrence of large forecast errors. A seemingly unnoticed phenomenon is that although the use of the L_1-loss in forecast combination often improves over the L_2-loss in obtaining more accurate forecast combinations, it may have a higher tendency to have large forecast errors. Therefore, unfortunately, as will be seen, a robust combining method may

actually work against the goal of having fewer outliers in the context of forecast combination.

In this chapter, we propose a synthetic loss function (denoted by the L_{210}-loss) which is a linear combination of the L_2-loss, the L_1-loss and a smoothed L_0-loss that naturally and smoothly penalizes the occurrence of large forecast errors more directly. It is used to propose a new combination algorithm based on the general AFTER scheme from Yang (2004). We establish oracle inequalities in terms of the L_{210}-loss that show optimal converging properties of the new AFTER method. Numeric results also support the advantages of our outlier-protective approach in terms of reducing the frequency of large forecast errors in the combined forecasts while maintaining comparable accuracy under both the L_2- and L_1-losses.

It should be pointed out that outlier forecasts can be defined in different ways, e.g., in relation to other candidate forecasts or to the observed value. In this work, an outlier forecast refers to a forecast that is far away from the realized value (i.e., the forecast error is large in absolute values). Forecasts that are drastically different from the majority in a panel of forecasts may also be defined as outliers. Such outliers may or may not be a concern in terms of forecast accuracy.

The plan of this chapter is as follows: section 2.2 discusses the motivation and the design of the loss function L_{210} with numeric examples demonstrating its efficiency in terms of outlier protection. In section 2.3, the L_{210}-loss based AFTER methods are proposed and theoretically examined. Simulation results are presented to evaluate the performance of our new combination approach in section 2.4. Real data from the M3-Competition (see e.g., Makridakis & Hibon, 2000) are used in section 2.5 and the results also confirm advantages of our methods. Section 2.6 concludes the chapter. The proofs of the theoretical results are presented in the appendix.

2.2 Outlier Protective Loss Functions

2.2.1 A Deficiency of the Robust L_1-loss

The L_1-loss is relatively more resistant to occasional outliers. This well-known nice feature is exploited in e.g., Wei & Yang (2012) for robust forecast combination, which results in more accurate forecasts. However, the robustness comes with a price: the L_1-loss is often less outlier protective in the sense that when used to compare different forecasts, it may not dislike enough forecasters that have higher frequency of outliers but with comparable (or slightly better) cumulative L_1-loss because it puts relatively less penalty (compared to e.g., the L_2-loss) to large forecast errors (outliers). For an understanding of this matter, examples will be provided after reviewing a framework to compare loss functions.

Objective Comparison of Loss Functions

The comparison of loss functions is usually entangled with the evaluation criteria used to define better forecasters, which typically involves loss functions. To avoid the difficulty due to the circular reference, Chen & Yang (2004) proposed a methodology to compare loss functions objectively.

In a time series setting, suppose we have a variable Y with two competing forecasters $\hat{Y}_1$ and $\hat{Y}_2$. Specifically, $\hat{Y}_{1,i}$ and $\hat{Y}_{2,i}$ are the forecasts for Y_i made at time $i-1$. Let $e_{1,i} = Y_i - \hat{Y}_{1,i}$ and $e_{2,i} = Y_i - \hat{Y}_{2,i}$ be the forecast errors. Suppose $e_{1,i}$ and $e_{2,i}$ are *iid* from certain distributions respectively, and let F_1 and F_2 be the cumulative distribution functions of $|e_{1,i}|$ and $|e_{2,i}|$ respectively.

If $F_1(x) \geq F_2(x)$ for all $x \geq 0$ (i.e., F_1 is stochastically smaller than or equal to F_2), then, theoretically, $E[L(|e_{1,i}|)] \leq E[L(|e_{2,i}|)]$ holds for any non-decreasing loss function L with $L(0) = 0$.

Therefore, theoretically, $\hat{Y}_1$ is a better forecaster than $\hat{Y}_2$ regardless of the loss

functions used for performance evaluation. However, different loss functions have different capabilities to pick out the better one. For example, if $e_{1,i}$ and $e_{2,i}$ are from $N(0,1)$ and $N(0,1.1^2)$ respectively, we generate samples $\{e_{1,i}\}_{i=1}^n$ and $\{e_{2,i}\}_{i=1}^n$ with size $n = 100$ independently for 10^8 times, then the cumulative L_2-loss has 83.4% chance to pick out $\hat{Y}_1$ (i.e., $\sum_{i=1}^n e_{1,i}^2 < \sum_{i=1}^n e_{2,i}^2$) in contrast to 81.3% for the L_1-loss (i.e., $\sum_{i=1}^n |e_{1,i}| < \sum_{i=1}^n |e_{2,i}|$). So, following this idea, by supplying two sequences of stochastically ordered errors (in absolute values), the one that is more likely to pick out the better forecaster should be considered the better loss function in a pair of competing loss functions. Thus, we can compare different loss functions objectively in a sensible aspect.

Example 1

In this example, in the same context described above, consider $e_{1,i}$ having 95% chance to follow $N(0,1)$ and 5% chance to follow a t_3-distribution (denoted by $95\%N(0,1) \oplus 5\%t_3$), $e_{2,i}$ follows the distribution of $1.05e_{1,i}$ and the sample size n is taken to be 30, 60, 100 and 200. A forecast error is considered to be large if its absolute value is larger than 2 in this example.

In this simulation, in Table 2.1, we present the probabilities that the L_2-loss (column 6) and the L_1-loss (column 7) (the L_{210}-loss will be defined later) pick out $\hat{Y}_1$ (the theoretically better one). An entry in column 2 is the (simulated) probability that the forecaster with smaller L_2-loss also has fewer large forecast errors. The same probabilities for the L_1-loss are in column 3.

From the comparison of columns 2 and 3, we see that the L_2-loss is more capable of picking out the forecaster with fewer outliers, while from columns 6 and 7, the L_1-loss is relatively more capable of identifying the better forecaster. The example reveals that the advantage of the L_1-loss in resisting the influence of outliers goes hand-in-hand with its disadvantage of being more likely to prefer the ones with more

Table 2.1: Performance evaluation criteria comparison (Example 1/Scenario 1)

	Outlier-protection				Choosing $\hat{Y}_1$			
	95% N(0,1) $\oplus$ 5% t_3							
n	L_2	L_1	$L_{210}^{(1)}$	$L_{210}^{(2)}$	L_2	L_1	$L_{210}^{(1)}$	$L_{210}^{(2)}$
30	0.759	0.711	0.756	0.799	0.678	0.680	0.678	0.675
60	0.778	0.740	0.779	0.830	0.736	0.739	0.744	0.750
100	0.794	0.769	0.808	0.846	0.779	0.798	0.800	0.796
200	0.836	0.832	0.857	0.879	0.848	0.880	0.884	0.878

outliers.

The average differences of the above probabilities between the L_2-loss and the L_1-loss are between 1-5% (with standard errors smaller than 10^{-4}), which is not necessarily practically insignificant. Note that differences between competing forecasting methods are often around 1-2% under various evaluation criteria (see e.g., Makridakis & Hibon, 2000).

2.2.2 L_{210}-loss

Since the L_1-loss takes care of the robustness efficiently and the L_2-loss is relatively more sensitive to (occasional) large forecast errors, a nature candidate to have a simultaneous control of both the robustness and outlier-protection tendency is a linear combination of the L_2- and L_1-losses as follows:

$$L_{21}(x|\alpha) = |x| + \alpha\frac{x^2}{m}, \tag{2.1}$$

where m is the median (or at that scale) of the absolute forecast errors, and α is a positive constant.

However, both the L_2- and L_1-losses (thus the L_{21}-loss) put indirect attentions to the occurrence of large forecast errors. To deal with the concern of large forecast errors upfront, for $0 < \gamma_1 \leq +\infty$ and $-\infty \leq \gamma_2 < 0$, we define the L_0-loss as:

$$L_0(e|\gamma_1, \gamma_2) = I(e \geq \gamma_1 \text{ or } e \leq \gamma_2). \tag{2.2}$$

It can be added to the L_{21}-loss in expression (2.1) to put more direct and significant penalty to the occurrence of large errors. The new synthetic loss function is denoted as L_{210}.

Obviously, the L_{210}-loss is not continuous (since the L_0-loss is generally not continuous). But continuity/smoothness is important for efficient computation when the loss is used to fit a model by empirical risk minimization (see, e.g., Liu & Wu, 2007), and it is also useful to the development of our theoretical results (as seen in the proofs of the theorems in this paper). So, a continuous surrogate of the L_0-loss in expression (2.2) can be a better alternative. In order to narrow down the choices, two constraints are considered:

1. The continuous surrogate should be close to the original L_0-loss;

2. The concavity from the surrogate L_0-loss function is not too large since the overall convexity of the corresponding L_{210}-loss function is very useful for numeric optimization and our theoretic development.

We choose a surrogate function in the form of the **M**inimax **C**oncavity **P**enalty (**MCP**) from Zhang (2010). Specifically, for the L_0-loss in (2.2), the **MCP** surrogate

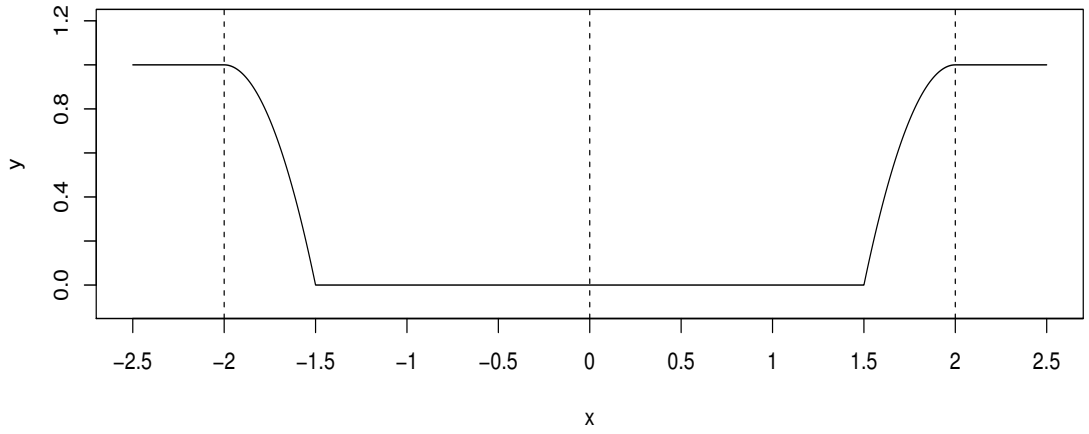

Figure 2.1: MCP surrogate of the L_0-loss

$\tilde{L}_0$-loss is:

$$\tilde{L}_0(e|\gamma_1,\gamma_2) = \begin{cases} 1, & \text{if } e \geq \gamma_1 \text{ or } e \leq \gamma_2 \\ 1 - \frac{1}{\gamma_1^2(1-r_1)^2}(e-\gamma_1)^2, & \text{if } r_1\gamma_1 \leq e \leq \gamma_1 \\ 1 - \frac{1}{\gamma_2^2(1-r_2)^2}(e-\gamma_2)^2, & \text{if } \gamma_2 \leq e \leq r_2\gamma_2 \\ 0, & \text{if } \gamma_2 r_2 \leq e \leq \gamma_1 r_1, \end{cases} \tag{2.3}$$

where $0 < r_1, r_2 < 1$, and they control how sharp the jumps from 0 to 1 are (the larger the sharper). This function has second derivative everywhere except at $e = \gamma_1 r_1$ and $e = \gamma_2 r_2$.

Figure 2.1 is an example of $\tilde{L}_0(e|\gamma_1 = 2, \gamma_2 = -2, r_1 = r_2 = 0.75)$.

Therefore, the continuous L_{210}-loss function we proposed is:

$$L_{210}(e) := |e| + \alpha_1 \frac{e^2}{m} + \alpha_2 m \tilde{L}_0(e|\gamma_1 m, \gamma_2 m, r_1, r_2), \tag{2.4}$$

where $\alpha_1 > 0$ and $\alpha_2 \geq 0$ are two constants. The choice of m is discussed in the first remark below. Also, an example of the specification of m in real data applications is given in section 2.5. Note also that asymmetric quadratic and absolute functions can be used instead of e^2 and $|e|$, respectively.

Remarks:

1. The use of m in the L_{210}-loss makes its three components at the same scale. One can choose m based on previous experience or the data at hand only. Of course, when one has strong evidence that the data generating process has changed, updating m based on the new information is necessary. Our numerical experience seems to suggest that in real application, choosing an m that is of the same scale of and not too far way from the median of the absolute forecast errors works well as seen in Scenarios 1 and 2 in the next subsection.

2. For α_1, it determines the degree of concern about the forecast errors under the L_2-loss. We need to point out that if the frequency of the large forecast errors is high rather than occasional, then the L_2-loss may become less sensitive to large forecast errors since the earlier large ones may dominate the whole cumulative loss quickly. A relatively small α_1, such as 0.5 or 0.1, is recommended if one does not have specific preferences.

3. The coefficient α_2 controls how much penalty the user wants to put on the occurrence of large errors. When the outlier-protection is of great importance, a larger α_2 may be explored.

4. The best choice of γ_1, γ_2, r_1 and r_2 may be case dependent. If there is no specific consideration for the parameters, $\gamma_1 = -\gamma_2 = 2$ and $0.5 \leq r_1 = r_2 \leq 0.9$ is a good starting point suggested by our numeric work.

2.2.3 L_{210}-loss as a Performance Evaluation Criterion

In this subsection, we show that using the L_{210}-loss leads to a more protective choice of a forecaster in terms of the frequency of outliers than that of the L_2- and L_1-losses. That is, given a pair of competing forecasters for $\{Y_i\}_{i=1}^n$, the L_{210}-loss is more likely to prefer the forecaster with fewer large forecast errors than the L_2- and L_1-losses.

Also, we show that the capability of the L_{210}-loss to identify the (theoretical) better forecaster is comparable to the better one of the L_2- and L_1-losses.

Scenario 1

Using the scenario in section 2.2.1, the $L_{210}^{(1)}$-loss and the $L_{210}^{(2)}$-loss are defined with common parameters (in expression (2.4)): $m = 1, \alpha_1 = 1, \alpha_2 = 3, \gamma_1 = 2, \gamma_2 = -2$. But the $L_{210}^{(1)}$-loss takes $r_1 = r_2 = 0.75$, and the $L_{210}^{(2)}$-loss takes $r_1 = r_2 = 0.9$. The results are in Table 2.1.

For the L_{210}-loss, from the comparison between columns 4 and 5, we see that its ability to pick out the forecaster with fewer large forecast errors gets better when the jump from 0 to 1 in the $\tilde{L}_0$-loss gets sharper. From columns 8 and 9, its capability to identify the better forecaster is slightly limited by the sharpness of the jump. Note that both the m in the two L_{210}'s are 1, which is not exactly equal but close to the theoretical medians of the absolute errors, and it works well (in other scenarios we tried as well). Also, other choices for the parameters in the L_{210}-loss are tried and similar stories are found.

Scenario 2

It is possible that the concern about the forecast outliers is not symmetric in the positive and negative directions. For this situation, an asymmetric L_{210}-loss can be defined with an asymmetric continuous surrogate of the L_0-loss. Below is an example

Table 2.2: Performance evaluation criteria comparison (Scenario 2)

	Outlier-protection					Capability to pick $\hat{Y}_1$		
n	L_2	L_1	$L_{210}^{(1)}$	$L_{210}^{(2)}$	$L_{210}^{(3)}$	$L_{210}^{(1)}$	$L_{210}^{(2)}$	$L_{210}^{(3)}$
30	0.674	0.643	0.760	0.774	0.704	0.576	0.580	0.559
60	0.663	0.636	0.768	0.792	0.714	0.596	0.619	0.601
100	0.638	0.620	0.767	0.793	0.721	0.630	0.664	0.628
200	0.601	0.589	0.777	0.813	0.744	0.685	0.721	0.685

for the efficiency of the asymmetric L_{210}-loss.

Using the notation from example 1 (section 2.2.1), let $e_{1,i} \sim 80\%N(0,1) \oplus 20\%(2-\Gamma(2,1))$ and $e_{2,i} \sim 80\%N(0,1) \oplus 20\%(\Gamma(2,1)-2)$, where $\Gamma(2,1)$ denotes the Gamma-distribution with shape parameter 2 and scale parameter 1. So, $E(e_{1,i}) = E(e_{2,i}) = 0$ and $e_{1,i}$ and $e_{2,i}$ are not symmetric about 0. If our concern is the frequency of the errors larger than 2 ($L_0(x) := I(x > 2)$), then, theoretically, forecaster $\hat{Y}_1$ is better than $\hat{Y}_2$. In this simulation, everything else remains the same as that in section 2.1.2.

The results at various sample sizes are summarized in Table 2.2. In Table 2.2, the $L_{210}^{(1)}$, $L_{210}^{(2)}$ and $L_{210}^{(3)}$ are defined with $m = 1$, $\alpha_1 = 1$,$\alpha_2 = 3$, $\gamma_2 = -\infty$, $r_1 = r_2 = 0.8$. For γ_1, it equals $2, 2.5$ and 3 in the $L_{210}^{(1)}$-, $L_{210}^{(2)}$- and $L_{210}^{(3)}$-losses respectively. From Table 2.2, we can see that:

1. The capacities of the L_2- and L_1-losses to pick out $\hat{Y}_1$ are omitted since they simply cannot tell the difference between $e_{1,i}$ and $e_{2,i}$.

2. By the help of the asymmetric $\tilde{L}_0$-loss, the L_{210}-loss is capable of capturing the asymmetric outliers. Further, from the results (some are not presented), the performance of the L_{210}-loss is not too sensitive to the choice of the parameters in the $\tilde{L}_0$-loss.

3. As the sample size increases, the advantages of the L_{210}-loss get more significant in terms of the capabilities to pick out the better forecasters and also the forecasters with fewer large errors (larger than 2).

2.3 Combination with Outlier Protective Loss Function

2.3.1 L_{210}-AFTER

Suppose we have N candidate forecasters, and combination starts at time n_0 (the first $n_0 - 1$ observations are used as training data). Let $\hat{y}_{j,i}$ be the forecast of y_i from candidate forecaster j. Accordingly, let $W_{j,i}$ be the combination weight of candidate j for y_i that satisfies $\sum_{j=1}^{N} W_{j,i} = 1$, and we start with $W_{j,n_0} = 1/N$. Let μ_i be the conditional mean of y_i given z^{i-1} (z^{i-1} represents the information available before observing y_i) and $e_i := y_i - \mu_i$.

Then, for $t \geq n_0 + 1$, the L_{210}-loss based AFTER weighting is:

$$W_{j,t} = \frac{\prod_{k=n_0}^{t-1} \exp\Big(-\lambda L_{210}(y_k - \hat{y}_{j,k})\Big)}{\sum_{j'=1}^{N} \prod_{k=n_0}^{t-1} \exp\Big(-\lambda L_{210}(y_k - \hat{y}_{j',k})\Big)}, \tag{2.5}$$

where λ is a positive constant that will be discussed later in this section. The combined forecast for y_t is defined as:

$$\hat{y}_t^* = \sum_{j=1}^{N} W_{j,t} \hat{y}_{j,t}. \tag{2.6}$$

In order to achieve a theoretical risk bound for this L_{210}-AFTER method, two

conditions are needed.

Condition 1: There exists a constant $\tau > 0$ such that $P(\sup_{i,j} |\hat{y}_{j,i} - \mu_i| < \tau) = 1$.

Condition 2: There exists a constant $s_0 > 0$ and two continuous functions $0 < H_1(s), H_2(s) < \infty$ on $(-s_0, s_0)$, such that $E_i \exp(s|e_i|) \leq H_1(s)$ and $E_i e_i^2 \exp(s|e_i|) \leq H_2(s)$ for all $s \in (-s_0, s_0)$ and all $i \geq n_0$ with probability 1, where E_i is the expectation conditional on z^{i-1}.

Theorem 1

Under Conditions 1 and 2, with a small enough positive constant λ, if the parameters of the L_{210}-loss function satisfy $\frac{\alpha_2}{\alpha_1} < \min\{\gamma_2^2(1-r_2)^2, \gamma_1^2(1-r_1)^2\}$, then

$$\frac{1}{n}\sum_{i=n_0}^{n+n_0-1} E\big[L_{210}(y_i - \hat{y}_i^*)\big] \leq \inf_{1\leq j\leq N}\Bigg(\frac{\log(N)}{\lambda n} + \frac{1}{n}\sum_{i=n_0}^{n+n_0-1} E\big[L_{210}(y_i - \hat{y}_{j,i})\big]\Bigg). \qquad \square$$

Remarks:

1. The theorem suggests that the combined forecast performs as well as the best individual candidate forecaster up to any given time plus a small penalty which decreases when the length of the evaluation periods gets larger.

2. The parameter λ in Theorem 1 depends on τ in Condition 1, s_0, H_1 and H_2 in Condition 2 and the parameters of the L_{210}-loss function.

3. Condition 1 simply requires that all the candidate forecasts are not too far away from the conditional means. It does not put any constraints on the boundedness of y (and thus allows severe outliers), and it certainly holds if the forecasts and the observations are bounded (which may be reasonable for many real applications), though theoretically it does not hold for some time series models

(such as AR(1); see Wei & Yang, 2012, for more discussion).

4. Condition 2 assumes that the error distribution in the true model does not have a tail that is heavier than an exponential-decay, which is satisfied by e.g. sub-Gaussian and double-exponential distributions.

5. The constraint of the parameters in the L_{210}-loss implies that the L_{210}-loss function is lower-bounded by a quadratic curve which we will use in the proof of Theorem 1. Also, it suggests that the penalty to the occurrence of large forecast errors can not be too large.

6. The combined forecast from the L_{210}-AFTER also provides a multi-objective combination which serves three evaluation criteria simultaneously: the L_2-, L_1- and L_0-losses.

The proof of Theorem 1 is available in the section 2.7.1.

2.3.2 Data Driven L_{210}-AFTER

The choice of the parameter λ in the weighting formula of expression (2.5) is a difficult issue since it depends on some unknown quantities as discussed in section 2.3.1. In this subsection, we propose a data-driven L_{210}-AFTER method that avoids this difficulty, and is thus more applicable in real situations. Before the introduction of the data-driven L_{210}-AFTER, a new distribution family is considered.

$\mathcal{F}_{210}$-Family

From Chen & Yang (2004), the L_2-loss based AFTER (L_2-AFTER) works efficiently for the Gaussian (or close to Gaussian) errors since the L_2-loss is the exponential kernel of the univariate Gaussian family. In contrast, the L_1-loss based AFTER (L_1-AFTER) often works better when the error distributions have heavier tails. In the

same spirit, the L_{210}-loss is associated with a density that has the L_{210}-loss in the exponential kernel.

We define a density family, called $\mathcal{F}_{210}$-family, that is associated with the L_{210}-loss. The probability density functions in this family are in the form

$$f(x|\delta) := \frac{1}{h(\delta)} \exp(-L_{210}(x)/\delta),$$

where $\delta > 0$ is a scale-parameter and $\frac{1}{h(\delta)}$ is a function of δ that normalizes $g(x|\delta) := \exp(-L_{210}(x)/\delta)$ to be a probability density function.

Note that the Gaussian or the double-exponential family can be considered as special cases in the $\mathcal{F}_{210}$-family, and the $\mathcal{F}_{210}$-family is more efficient in describing the error distributions with more likely occurrence of outliers.

In the following subsection, we present a version of the L_{210}-AFTER with estimation of δ to avoid the difficulty in specifying λ. The related numeric experiments are provided in section 2.4.

L_{210}-AFTER with Scale-parameter Estimation

The new weighting formula is:

$$W_{j,t} = \frac{\prod_{k=n_0}^{t-1} \frac{1}{\sqrt{\hat{\delta}_{j,k}}} \exp\left(-L_{210}(y_k - \hat{y}_{j,k})/\hat{\delta}_{j,k}\right)}{\sum_{j'=1}^{N} \prod_{k=n_0}^{t-1} \frac{1}{\sqrt{\hat{\delta}_{j',k}}} \exp\left(-L_{210}(y_k - \hat{y}_{j',k})/\hat{\delta}_{j',k}\right)}, \tag{2.7}$$

where $\hat{\delta}_{j,k}$ is an estimate of δ_k (the conditional scale-parameter given z^{k-1}) from forecaster j at time period $k-1$ (an example choice to estimate $\hat{\delta}_{j,k}$ is in Remark 3 after Theorem 2). The combined forecast for y_t is the same as in expression (2.6).

Besides point forecast of y_t, prediction of the whole distribution of y_t ($t \geq n_0 + 1$)

conditional on z^{t-1} is often of interest (see, e.g., Timmermann, 2000; Yang, 2000). With the weights $W_{j,t}$, a nature forecast of the conditional distribution of y_t (denoted as q_t and $q_t = \frac{1}{h(\delta_t)} \exp(-L_{210}(y_t - \mu_t)/\delta_t)$) is

$$\hat{q}_t = \sum_{j=1}^{N} W_{j,t} \frac{1}{h(\hat{\delta}_{j,t})} \exp(-L_{210}(y_t - \hat{y}_{j,t})/\hat{\delta}_{j,t}).$$

It is well know that Kullback-Leibler divergence is a proper measure of the distance between two densities. Let $D(q_t||\hat{q}_t)$ denotes the K-L divergence between q_t and $\hat{q}_t$ (conditional on z^{t-1}). Then the expectation of $\frac{1}{n}\sum_{t=n_0}^{n_0+n-1} D(q_t||\hat{q}_t)$ is a natural measure of the overall performance of $\hat{q}_t$ over time.

Condition 3: There exists a constant $A \geq 1$ such that $1/A \leq \delta_i, \hat{\delta}_{j,i} \leq A$ for all i, j with probability 1.

Theorem 2

Let $y_i = \eta_i + \epsilon_i$, where ϵ_i follows a distribution from the F_{210}-family with unknown scale-parameter δ_i. Under Condition 3, we have

$$\frac{1}{n}\sum_{t=n_0}^{n_0+n-1} ED(q_t||\hat{q}_t) \leq \inf_{1\leq j\leq N}\Bigg(\frac{\log(N)}{n} + \frac{C}{n}\sum_{i=n_0}^{n_0+n-1}\Big(E|L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)| + E|\hat{\delta}_{j,i} - \delta_i|\Big)\Bigg),$$

where C is a constant that depends on A and the parameters in expression (2.4). □

Theorem 2 states that the average risk of the combined forecast is bounded in order by the averaged mean absolute differences between the L_{210}-loss of the combined forecast and the L_{210}-loss of η_i's plus two additional terms, namely, the estimation accuracy for δ's and the log size of the candidate pool relative to the sample size n. Remarks:

1. The newer version of the L_{210}-AFTER in Theorem 2 has less restriction on the coefficient parameters α_1 and α_2, the thresholds γ_1 and γ_2, and steepness parameters r_1 and r_2 in defining the L_{210}-loss. For example, it is now allowed to put a very large α_2 to reflect a strong dislike of occurrence of the large forecast errors without invalidating the theoretical property in the theorem.

2. Condition 3 constrains the scale parameters and their estimators to be in a compact set away from zero and infinity.

3. A natural choice for $\hat{\delta}_{j,k}$ is that $\hat{\delta}_{j,k} := \frac{1}{k-1}\sum_{l=1}^{k-1} L_{210}(y_l - \hat{y}_{j,l})$. This is our choice for the numeric examples in the following sections.

The proof of Theorem 2 is provided in the section 2.7.2.

2.4 Simulation Results

In this section, simulation results are presented to demonstrate advantages of the L_{210}-AFTER. In this and the next sections, the L_{210}-AFTER refers to the data-driven version, and the L_2- and L_1-AFTERs refer to the versions in sections 2 and 3.2 of Wei & Yang (2012), respectively.

In the general expression of the L_{210}-loss, there are several parameters, among which γ_1 and γ_2 can be determined by the interests of the specific applications and r_1 and r_2 control the approximations to the L_0-loss by the smooth surrogate. The parameters α_1 and α_2 are the least guided. To have an informative but focused study, in this and the next sections, unless otherwise stated, we use $\gamma_1 = 2$, $\gamma_2 = -2$, and $r_1 = r_2 = 0.9$ in the L_{210}-AFTERs and consider the loss function $L_0(e) = I(|e| > 2m)$. In the simulations, m is the median of the absolute value of the innovation error. In

all settings, multiple choices of α_1 and α_2 are investigated systematically. In addition, asymmetric $\tilde{L}_0$ component in the L_{210}-loss is considered in some cases.

2.4.1 Simulation Setup

The candidate forecasts are generated by linear regression models. The possible large forecast errors are designed to come from the innovation errors.

In all the settings below, we have 5 predictors, $X_1, \cdots, X_5$, and they are randomly generated from certain distributions (to be specified). The true model is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p_0} X_{p_0} + \epsilon, \tag{2.8}$$

where $1 \leq p_0 \leq 5$ and ϵ is generated from a certain distribution.

The forecast candidates are obtained from the linear regression models as follows: $Y = \beta_0 + \beta_1 X_1 + e$, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, $\cdots$, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + e$. Least squares estimates are used as the estimates of the parameters in each model, based on which the forecasts are made.

The detailed simulation procedure is:

Step 1: Generate $\beta = (\beta_1, \cdots, \beta_{p_0})$ in expression (2.8);

Step 2: Generate 150 iid copies of $\{X_1, \cdots, X_5\}$ and ϵ;

Step 3: Generate 150 Y values based on the expression ((2.8)) using the β from Step 1, the $\{X_1, \cdots, X_5\}$ and ϵ from Step 2;

Step 4: With the 150 observations of $\{X_1, \cdots, X_5, Y\}$ generated from Steps 2 and 3, in a sequential fashion, after the 30-th observation, the candidate forecasts (from the aforementioned 5 models) are obtained for the different time periods. For each combination method, the first 10 forecasting periods are used as training and the L_2-, L_1- and L_0-losses are calculated beginning at the 41st observation, i.e., the cumulative loss for the j-th forecaster is $\sum_{t=41}^{150} L(\mu_t - \hat{y}_{j,t})$, where L is one of the three losses.

Note that, since we know the true means of the observations in simulations, combined forecasts are compared with the true means under loss function L to have a better comparison;

Step 5: Repeat Steps 2-4 200 times independently and record the averaged L_2-, L_1- and L_0-losses (over the 200 replications) for each combination method;

Step 6: For the averaged L_2- and L_1-losses from Step 5, ratios of the losses of other methods over that of the L_1-AFTER are recorded. For the averaged L_0-loss from Step 5, the differences (other methods minus that of the L_1-AFTER) are recorded;

Step 7: Repeat Steps 1-6 M times independently (see the specific choice of M in the description of each scenario below), and the summaries (mean, standard error and median) over the M sets of ratios and differences are presented.

2.4.2 The Competing Forecast Combination Methods Considered

We intend to compare the performances of the L_{210}-AFTER with several popular forecast combination methods, including simple average (SA), trimmed mean (TM), median (MD), variance-covariance estimation based combination (BG), combination via linear regression (LR) and constrained linear regression (CLR) and the existing AFTER methods.

2.4.3 Scenarios

Scenario 3

In this scenario, $\{X_1, \cdots, X_5\}$ are from a Normal distribution with zero mean and covariance matrix Σ with entry $\Sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 5$. The p_0 in expression (2.8) is randomly picked from $\{1, 2, \cdots, 5\}$ with equal probabilities. That is, in each repeat of the Steps 1-6 of the Step 7, we first generate a p_0 and then generate a set

of β with size p_0 in the Step 1. M in Step 7 is 10^6. A large M is used here because we want to show the differences (and how stable they are) among some of the L_{210}-AFTERs which have very close performances. Let ϵ have a mixture distribution with probability 90% from $N(0,1)$ and 10% from $Unif[-5,5]$, and the components of β be iid from $Unif[-1,1]$.

The simulation results of $m = 0.8$ (not equal but close to the median of the absolute value of the error) are summarized into Table 2.3, Table 2.4 and Table 2.5. Results with some other choices of the parameters, e.g., $m = 0.6$, $m = 1$, $r_1 = r_2 = 0.75$, and Σ with $\Sigma_{i,j} = I(i = j)$ for $1 \leq i, j \leq 5$ are not included because they provide basically the same stories.

Note that, in Table 2.3, the values above the parenthesis are the means of the ratios or differences relative to the L_1-AFTER. The values in the parenthesis are the medians. The standard errors for the means of the L_2-AFTER and the LRC are smaller than 4×10^{-4} and smaller than 10^{-3} for other methods. In Table 2.4, the columns 2-5 (6-9) are the mean of the L_2-loss (L_1-loss) ratios of the L_{210}-AFTER over the L_1-AFTER.

In Table 2.5, the columns 2-5 are the differences of the number of outliers from the L_{210}-AFTER compared to the L_1-AFTER. The last 4 columns are the probabilities that the related L_{210}-AFTER provide forecasts with fewer outliers than that of the L_1-AFTER. The standard errors for all the values are smaller than 10^{-3}.

Scenario 4

Consider an asymmetric modification of Scenario 3. Let ϵ now be from a mixture distribution with probability 90% from $N(-0.4,1)$ and 10% from $Unif[2,5.2]$, which has mean zero but with more likelihood to have positively large forecast errors. The only other change is to use an asymmetric loss $L_0(e) = I(e > 2m)$ and the corresponding $\gamma_1 = 2$ and $\gamma_2 = -\infty$. The results are summarized into Table 2.6, Table 2.7

and Table 2.8 and they are organized in the same ways as those for Scenario 3. Note that the standard errors for all the values in Table 2.6 and Table 2.8 are smaller than 10^{-3}.

2.4.4 Results

The simulation results are summarized into tables in this and the following sections and on these tables, the L_2- and L_1-AFTERs are denoted as L_1A and L_2A, respectively.

The Comparison Inside the AFTER Family

Since the L_{210}-AFTER is designed to provide extra outlier-protection over the existing AFTER methods, we compare it with the L_2- and L_1-AFTERs first.

1. For Scenarios 3 and 4 (Table 2.3-Table 2.8) , we see that the overall performance (under the L_1- and L_2-losses) of the L_{210}-AFTER is comparable to that of the L_2- and L_1-AFTERs, while the L_{210}-AFTER is more efficient in terms of outlier protection (under the L_0-loss). In fact, properly selected $\{\alpha_1, \alpha_2\}$ may even enable the L_{210}-AFTER to outperform the L_2- and L_1-AFTERs under all the three loss functions sometimes.

2. From the results, we see that the L_{210}-AFTER is more outlier-protective than the L_1-AFTER. The differences of the numbers of outliers (defined by the L_0-loss) are about -0.1 or -0.2, which is non-trivial since the average number of outliers is about 1.5-2.5 for both cases in the evaluation periods of the candidate forecasts.

3. For some set of $\{\alpha_1, \alpha_2\}$, the L_{210}-AFTER fails to improve over the L_2-AFTER in terms of outlier protection. This suggests that the selection/tuning of the pa-

rameters in the L_{210}-AFTER should not be done carelessly. A general guideline of choosing the parameters efficiently is presented in section 2.4.3.

4. We have also considered other error distributions, such as t_4 or mixture distributions with different mixing probabilities. The relative performances between L_1-AFTER and L_2-AFTER can be different, but the relative behavior of the L_{210}-AFTER is quite consistent, although in some cases its benefit is less visible.

The L_{210}-AFTER vs. Other Methods

Here, we compare the L_{210}-AFTER with other popular combination methods.

1. Overall, from Table 2.3-Table 2.8, the L_{210}-AFTER outperforms all other competing methods outside the AFTER family under the L_2-, L_1- and L_0-loss functions.

2. The LRC is the best method outside the AFTER family. But in terms of outlier protection, the LRC is outperformed by most versions of the L_{210}-AFTER.

Roles of α_1 and α_2 in the L_{210}-AFTER

From our investigations in sections 2.4.4, the value of $\{\alpha_1, \alpha_2\}$ in the L_{210}-AFTER does affect its performances. In real applications, to train/tune the parameters in the L_{210}-AFTER on a training data set for further use is a proper strategy. Some general guidance on choosing these parameters properly can be helpful.

Table 2.3-Table 2.8 provide a general and intuitive understanding of how to choose proper parameters in the L_{210}-AFTER.

1. In general, the performances of the L_{210}-AFTER is fairly robust since a wide range of α_1 and α_2 combination equipped L_{210}-AFTERs perform quite similarly.

Table 2.3: Popular combination methods under the L_2-, L_1- and L_0-losses (Scenario 3)

	L_2A	LR	CLR	SA	MD	TM	BG
L2	0.825	3.796	0.870	1.161	1.229	1.155	1.024
	(0.799)	(3.464)	(0.876)	(1.012)	(1.034)	(0.999)	(0.970)
L1	0.920	1.703	0.942	1.124	1.113	1.104	1.057
	(0.910)	(1.663)	(0.950)	(1.041)	(1.030)	(1.038)	(1.021)
L0	-0.280	2.899	-0.259	0.257	0.751	0.420	-0.068
	(-0.130)	(2.880)	(-0.080)	(-0.035)	(-0.010)	(-0.030)	(-0.070)

2. From Table 2.5 and Table 2.8 for the different options of $\{\alpha_1, \alpha_2\}$, we observe that when α_1 is not large, increasing α_2 in a certain range enhances the advantages of outlier protection. When α_1 gets larger, the enhancement becomes relatively less significant. Since a large α_1 may damage the performance under the L_1- or L_2-loss, a moderate α_1 and non-zero α_2 can provide a better balance of the performances under the L_0-, L_1- and L_2-losses.

3. It is certainly not true that a larger α_2 makes the L_{210}-AFTER more outlier protective because it may sacrifice the usual forecast accuracy too much and mess up with the goal. Fortunately, α_2 does not need to be very large to put enough emphasis on the protection over outliers. The results suggest that if we have historical data, we can start with a small α_2 and increase it gradually to search for a good choice for outlier protection while not losing much efficiency in the L_2- and L_1-losses.

Table 2.4: The L_{210}-AFTER under the L_2- and L_1-losses (Scenario 3)

	Under L_2-loss				Under L_1-loss			
$\alpha_2\backslash\alpha_1$	3	2	1	0.5	3	2	1	0.5
10	0.811	0.806	0.807	0.812	0.915	0.912	0.912	0.914
5	0.810	0.805	0.814	0.813	0.914	0.912	0.915	0.913
3	0.810	0.805	0.819	0.821	0.915	0.912	0.918	0.913
1	0.811	0.807	0.826	0.829	0.915	0.912	0.921	0.915
1/5	0.811	0.808	0.830	0.836	0.915	0.913	0.923	0.914
0	0.812	0.809	0.832	0.838	0.917	0.915	0.923	0.915

Note: The standard errors for all the ratios and percentages are smaller than 5×10^{-4}.

Table 2.5: The L_{210}-AFTER under the L_0-loss (Scenario 3)

	Under L_0-loss				Chances of beating L_1A			
$\alpha_2\backslash\alpha_1$	3	2	1	0.5	3	2	1	0.5
10	-0.302	-0.304	-0.290	-0.282	0.812	0.825	0.808	0.789
5	-0.304	-0.307	-0.294	-0.273	0.809	0.820	0.804	0.783
3	-0.305	-0.308	-0.288	-0.264	0.807	0.817	0.793	0.778
1	-0.303	-0.306	-0.280	-0.253	0.804	0.811	0.800	0.774
1/5	-0.301	-0.305	-0.276	-0.242	0.802	0.807	0.795	0.770
0	-0.301	-0.304	-0.274	-0.244	0.803	0.807	0.794	0.766

Table 2.6: Popular combination methods under the L_2-, L_1- and L_0-losses (Scenario 4)

	L_2A	LR	CLR	SA	MD	TM	BG
L2	0.843	3.936	0.887	0.992	1.062	1.004	0.935
	(0.837)	(3.588)	(0.902)	(0.913)	(0.962)	(0.884)	(0.862)
L1	0.926	1.717	0.944	1.032	1.039	1.025	1.000
	(0.919)	(1.688)	(0.947)	(1.001)	(0.982)	(0.961)	(0.963)
L0	-0.199	2.293	-0.169	-0.024	0.221	0.061	-0.109
	(-0.070)	(2.275)	(-0.030)	(-0.060)	(-0.020)	(-0.050)	(-0.080)

Table 2.7: The L_{210}-AFTER under the L_2- and L_1-losses (Scenario 4)

	Under L_2-loss				Under L_1-loss			
$\alpha_2\backslash\alpha_1$	3	2	1	0.5	3	2	1	0.5
10	0.833	0.828	0.840	0.897	0.922	0.919	0.925	0.953
5	0.832	0.827	0.844	0.912	0.922	0.919	0.927	0.960
3	0.833	0.827	0.847	0.919	0.922	0.919	0.928	0.963
1	0.833	0.828	0.850	0.925	0.922	0.919	0.930	0.966
1/5	0.833	0.828	0.852	0.928	0.922	0.919	0.931	0.967
0	0.833	0.828	0.852	0.929	0.922	0.920	0.931	0.967

Note: The standard errors for all the ratios and percentages are smaller than 5×10^{-4}.

Table 2.8: The L_{210}-AFTER under the L_0-loss (Scenario 4)

	Under L_0-loss				Chances of beating L_1A			
$\alpha_2\backslash\alpha_1$	3	2	1	0.5	3	2	1	0.5
10	-0.207	-0.213	-0.200	-0.154	0.875	0.881	0.838	0.750
5	-0.210	-0.214	-0.197	-0.138	0.875	0.875	0.831	0.744
3	-0.209	-0.214	-0.195	-0.130	0.875	0.881	0.825	0.719
1	-0.209	-0.213	-0.193	-0.121	0.875	0.888	0.825	0.716
1/5	-0.209	-0.212	-0.190	-0.117	0.875	0.888	0.812	0.712
0	-0.206	-0.208	-0.190	-0.116	0.875	0.881	0.806	0.706

2.5 Real Data Example

In this section, we use real data to study the performance of the L_{210}-AFTER and compare it with several other combination methods. Both symmetric and asymmetric L_0-loss functions are considered to define forecast outliers and the associated L_{210}-AFTERs are applied.

The M3-competition data are a collection of 3003 real time series from various fields (e.g., business, finance, and economy) and 24 forecasters made predictions for each variable. This data set has been widely used to compare the efficiency of different forecasting methods (see, e.g., Makridakis & Hibon, 2000; Armstrong, 2007).

There are three different horizons (length) of the forecasts: 6, 8 and 18. Note that the forecasts by the forecasters were made all at once (1-step ahead, 2-step ahead,..., up to 6-, 8- or 18-step ahead). We choose the ones with 18 forecasts (1428 out of 3003: N1402 to N2829) for two main reasons: 1). Some of the candidate competing methods need a few data points to train the parameters before achieving a reasonable reliability. For example, to estimate the conditional variances used in the BG, at least

3-5 previous forecast errors are needed. 2). In order to evaluate the performance of the methods more effectively, a reasonable number of forecast periods is required and usually the larger the better.

2.5.1 The Competing Combination Methods

Except the linear regression related combination strategies, all other methods used in Scenario 4 are considered. The reason we exclude them is because we have way more forecasters than the prediction periods.

2.5.2 The Procedures

The Performance Measures

We use the simple average strategy as the benchmark since it is one of the simplest methods with reasonable performances and of great popularity in application.

Three loss functions are considered to summarize the performance of each method on each variable. Under the L_2-loss (L_1-loss) function, the mean squared (absolute) forecast error of another method over that of the simple average strategy is recorded for each variable. The summaries (mean, standard error and median) of the ratios over the set of variables are provided. For the L_0-loss function, the number of large forecast errors of each combination method, which will be defined in the following subsection, minus that of the simple average strategy is recorded for each variable. The summaries of the differences are provided.

We first compare the performances of the methods over all the 1428 variables and the summaries are in Table 2.9 (under the symmetric L_0-loss) and Table 2.10 (under the asymmetric L_0-loss). Then, a more specific comparison is performed. Since the L_{210}-AFTER is proposed to have a better control of the occurrence of large forecast errors, it is especially meaningful to be applied when the L_1-AFTER (one of the best

methods in the general comparison) performs poorly in that regard. Thus we focus on the series that the L_1-AFTER fails to beat the simple average strategy in outlier-protection (under each of the two L_0-loss functions) to have a more comprehensive understanding of the performance of the L_{210}-AFTER. The results are summarized into Table 2.11 (under the symmetric L_0-loss) and Table 2.12 (under the asymmetric L_0-loss).

The Parameters in the L_{210}-AFTER

For each variable, the combination starts at the 5-th forecasts, and the evaluation starts after the 8-th combination.

The choice of m in the L_{210}-loss (thus the L_{210}-AFTER) is the median of the absolute forecast errors of all candidate forecasts on the first 4 forecast periods. For the symmetric L_0-loss case, a forecast error is considered to be large when its absolute value is greater than $6m$ (a smaller choice such as $2m$ would end up with too many large forecast errors due to the difficulty of forecasting in the M3 competition). So, accordingly, $(\gamma_1, \gamma_2) = (6, -6)$. Smaller values for (γ_1, γ_2), such as $(5, -5)$ and $(4, -4)$, are also considered and they support the advantages of the L_{210}-AFTER in terms of outlier protection as well. Other options of r_1 and r_2 than $r_1 = r_2 = 0.9$ are tried, with similar results.

For the α_1 and α_2 in the L_{210}-loss function, we provide the results of multiple options to show: 1) Even for the general suggestions of the α_1 and α_2 without knowing the details of the target problems, the performance of the L_{210}-AFTER is still competitive; 2) The performance of the L_{210}-AFTER is fairly robust since similar results are found for reasonable wide ranges of α_1 and α_2.

Further, since the main goal here is to show the advantages of the L_{210}-AFTER in outlier protection, we use a relatively small α_1 to make the role of the $\tilde{L}_0$ more visible. Specifically we consider $\alpha_1 \in \{0.15, 0.03\}$ and $\alpha_2 \in \{3, 0.15\}$.

For the asymmetric L_0-loss case, we have $L_0(e) = I(e > 6m)$ and $(\gamma_1, \gamma_2) = (6, -\infty)$ in the L_{210}-AFTER.

2.5.3 Results

Comparing Different Schemes on the 1428 Variables

Table 2.9 and Table 2.10 provide the the comparison among methods over the 1428 variables under the L_2-, L_1- and L_0-losses.

Table 2.9: Relative performance over the SA on the M3-competition Data (Symmetric case)

		TM	MD	BG	L_1A	L_2A	$L_{210}A$	$L_{210}A$	$L_{210}A$	$L_{210}A$
	α_1						0.15	0.15	0.03	0.03
	α_2						3	0.15	3	0.15
	Mean	0.990	1.048	0.783	0.717	0.702	0.887	0.880	0.845	0.853
L_2	Se	0.003	0.009	0.009	0.016	0.016	0.032	0.035	0.026	0.036
	Median	1.000	1.024	0.845	0.660	0.654	0.683	0.684	0.669	0.668
	Mean	0.992	1.013	0.851	0.770	0.765	0.825	0.823	0.812	0.811
L_1	Se	0.002	0.005	0.006	0.009	0.009	0.011	0.011	0.011	0.011
	Median	1.000	1.012	0.911	0.797	0.791	0.798	0.799	0.798	0.799
L_0	Mean	-0.007	0.021	-0.364	-0.543	- 0.550	-0.560	-0.562	-0.568	-0.576
	Se	0.010	0.018	0.034	0.044	0.045	0.046	0.046	0.047	0.046

Note: The medians of all the methods under the L_0-loss are zero.

Table 2.10: Relative performance over the SA on the M3-competition Data (Asymmetric case)

		TM	MD	BG	L_1A	L_2A	$L_{210}A$	$L_{210}A$	$L_{210}A$	$L_{210}A$
	α_1						0.15	0.15	0.03	0.03
	α_2						3	0.15	3	0.15
	Mean	0.990	1.048	0.783	0.717	0.702	0.886	0.880	0.842	0.853
L_2	Se	0.003	0.009	0.009	0.016	0.016	0.032	0.035	0.026	0.036
	Median	1.000	1.024	0.845	0.660	0.654	0.683	0.684	0.667	0.668
	Mean	0.992	1.013	0.851	0.770	0.765	0.824	0.822	0.811	0.811
L_1	Se	0.002	0.005	0.006	0.009	0.009	0.011	0.011	0.011	0.011
	Median	1.000	1.012	0.911	0.797	0.791	0.798	0.799	0.796	0.799
L_0	Mean	-0.005	0.000	-0.116	-0.160	-0.161	-0.146	-0.153	-0.158	-0.165
	Se	0.009	0.002	0.016	0.022	0.022	0.028	0.028	0.027	0.027

Note: The medians of all the methods under the L_0-loss are zero.

We can see that:

1. The overall performances of the AFTER methods on these 1428 variables are significantly better than the best of all the other combination methods under the three loss functions (both symmetric and asymmetric L_0-loss cases). For example, under the L_2-loss, the accuracy of the combined forecasts from the L_2-AFTER is about 10% better than that of the BG, which is the best of the methods outside the AFTER family.

2. The performance of the L_{210}-AFTER is fairly robust when α_1 and α_2 are chosen in our explored ranges. In fact, given α_1 or α_2, the change of the other parameter in a reasonable range does not change the performance of the L_{210}-AFTER that much.

The L_{210}-AFTER vs. the L_1- and L_2-AFTERs

Now, we focus on the ones where the L_1-AFTER fails to beat the SA in terms of outlier protection. In fact, on 22 out of the 1428 variables, the SA beats the L_1-AFTER under the symmetric L_0-loss function and under the asymmetric L_0-loss function, the SA beats the L_1-AFTER on 12 variables. The results are in Table 2.11 and Table 2.12, which are organized in the same way as Table 2.9 and Table 2.10.

Table 2.11: The L_{210}-AFTER vs. Other methods when the SA beats the L_1-AFTER under the symmetric L_0-loss

		TM	MD	BG	L_1A	L_2A	$L_{210}A$	$L_{210}A$	$L_{210}A$	$L_{210}A$
	α_1						0.15	0.15	0.03	0.03
	α_2						3	0.15	3	0.15
	Mean	0.996	1.362	1.188	2.137	2.076	3.731	3.596	2.811	3.603
L_2	Se	0.024	0.129	0.108	0.559	0.591	1.623	1.918	1.080	1.916
	Median	1.008	1.165	1.081	1.519	1.493	1.073	1.043	1.114	1.114
	Mean	0.992	1.119	1.035	1.280	1.248	1.299	1.287	1.286	1.289
L_1	Se	0.011	0.047	0.038	0.098	0.091	0.129	0.129	0.119	0.130
	Median	0.991	1.049	1.020	1.166	1.159	1.051	1.071	1.057	1.061
L_0	Mean	0.001	1.136	0.500	1.682	1.591	1.000	0.909	0.864	0.682
	Se	0.066	0.035	0.109	0.232	0.204	0.240	0.242	0.259	0.210

Note: For the medians under the L_0-loss, the TM is 0, the MD is 0.5 and all other methods are 1.

Table 2.12: The L_{210}-AFTER vs. Other methods when the SA beats the L_1-AFTER under the Asymmetric L_0-loss

		TM	MD	BG	L_1A	L_2A	$L_{210}A$	$L_{210}A$	$L_{210}A$	$L_{210}A$
	α_1						0.15	0.15	0.03	0.03
	α_2						3	0.15	3	0.15
	Mean	1.537	0.997	1.164	2.538	2.433	1.991	1.689	1.865	1.791
L_2	Se	0.035	0.197	0.207	1.025	1.093	0.625	0.551	0.634	0.632
	Median	1.007	1.239	0.914	1.532	1.310	0.998	0.991	1.251	1.047
	Mean	0.999	1.199	1.018	1.346	1.287	1.170	1.127	1.190	1.168
L_1	Se	0.014	0.074	0.077	0.175	0.172	0.143	0.123	0.139	0.135
	Median	0.998	1.121	0.992	1.190	1.175	1.099	1.096	1.169	1.112
L_0	Mean	0.000	1.083	0.250	1.917	1.667	1.250	1.083	1.083	0.917
	Se	0.000	0.609	0.130	0.434	0.355	0.446	0.398	0.468	0.398

Note: For the medians under the L_0-loss, the TM, MD and BG is 0 and all other methods are 1.

1. Since we use the same set of parameters in the L_{210}-AFTER over all the variables, the performance of the L_{210}-AFTER may be limited. In spite of this, the results show that when the L_1-AFTER fails to control the presence of outliers effectively, the L_{210}-AFTER is a better option. Specifically, the L_{210}-AFTER provides about 0.6 - 1 fewer large forecast errors out of 10 evaluation periods on average.

2. The L_{210}-AFTER has comparable performance under the L_2- and L_1-losses to the L_1- and L_2-AFTERs.

3. The L_2- and L_{210}-AFTERs fail to beat the SA on these subsets of variables under all the three losses.

2.6 Conclusion

The choice of a loss function in forecast combination plays a very important role in constructing forecast combination weights. The quadratic loss (L_2-loss) has been the most commonly used. One major drawback is that the resulting combined weights may be overly influenced by a few outlier forecasts. The absolute loss (L_1-loss) leads to more robust weights, but on the other hand can actually perform worse in that its combined forecast may have a higher likelihood of producing outlier forecasts due to its downplaying the large errors than the quadratic loss, as seen in this work.

When even occasional outlier forecasts may have severe practical consequences, the new synthetic L_{210}-loss that directly addresses the concern can be used instead. When employed in the AFTER scheme, it is shown by simulations and real data to achieve the desired effect of reducing the occurrence of large forecast errors while maintaining forecast accuracy in the L_2- and L_1-losses. Oracle inequalities on forecast risks of the L_{210}-AFTER show that the combined forecasts or the associated density

estimates are close to the best candidates or the best density forecasts.

There are several parameters in the L_{210}-loss. The coefficients α_1 and α_2 decide the degree of emphasis on the L_2 and L_0 component, respectively, in the overall loss. The thresholds γ_1 and γ_2 indicate the largeness of the forecast error to be considered as an outlier. It is unlikely that one set of choices of these parameters works well generally. In this paper we have demonstrated numerically that our example choices performed quite satisfactorily in the presented settings. In real application, one can utilize subject knowledge or prior experience to have a synthetic loss that fits well the specific forecasting problem at hand.

2.7 Proofs of Theorems 1 and 2

2.7.1 Proof of Theorem 1

Since the maximum concavity of the $\tilde{L}_0$ in L_{210} is $\max\{\frac{\alpha_2}{m\gamma_2^2(1-r_2)^2}, \frac{\alpha_2}{m\gamma_1^2(1-r_1)^2}\}$. Thus the convexity of $L_{210}(x)$ holds when

$$\min\{2\alpha_1 - 2\alpha_2\gamma_1^2(1-r_1)^2, 2\alpha_1 - 2\alpha_2\gamma_2^2(1-r_2)^2\} \geq 0.$$

So, $\dfrac{\alpha_2}{\alpha_1} < \min\{\gamma_2^2(1-r_2)^2, \gamma_1^2(1-r_1)^2\}$ grantees that the function L_{210} (strongly) is convex (see, e.g., Nesterov, 2004, for more details).

Therefore, it is easy to see that for any a and $T > 0$, there exists $\bar{c} > 0$ and $\underline{c} > 0$ that:

$$\max_{-T\leq a\leq T} |L'_{210+}(a)| \leq \bar{c}(1+T), \quad \max_{-T\leq a\leq T} |L'_{210-}(a)| \leq \bar{c}(1+T),$$

and from the strong convexity of L_{210} that satisfies the condition given in Theorem

1, for a supporting hyperplane $y = \theta_{a_0}(a - a_0) + L_{210}(a_0)$ at any a_0, we have:

$$L_{210}(a) - (\theta_{a_0}(a - a_0) + L_{210}(a_0)) \geq \underline{c}(a - a_0)^2.$$

Then, define $h(x) = \exp(-\lambda L_{210}(x))$ and

$$q^n = \sum_{j=1}^{\infty} \frac{1}{N} \prod_{i=n_0}^{n_0+n-1} h(y_i - \hat{y}_{j,i}).$$

For any fixed j, we have $-\log(q^n) \leq \log(N) + \lambda \sum_{i=n_0}^{n_0+n-1} L_{210}(y_i - \hat{y}_{j,i})$.

By Lemma 10.1 of Catoni (1999) or Lemma 3.6.1 of Catoni (2004), under Condition 2, we have

$$\log(E^J \exp\{-\lambda L_{210}(y_i - \hat{y}_{J,i})\}) \leq -\lambda E^J L_{210}(y_i - \hat{y}_{J,i}) + I,$$

where

$$I = \frac{\lambda^2}{2} E^J \Big[L_{210}(Y_i - \hat{y}_{J,i}) - E^J \big[L_{210}(Y_i - \hat{y}_{J,i}) \big] \Big]^2 \times \exp\bigg(2\bar{c}\lambda \Big(|Y_i - \mu_i| + \big(1 + \sup_{j \geq 1} |\hat{y}_{j,i} - \mu_i| \big) \Big) \bigg),$$

and E^J denotes the expectation with respect to J with $P(J = j) = W_{j,i}$ for a fixed i.

Under Condition 2, let E_i denotes the conditional expectation given z^{i-1}, it follows, when $2\bar{c}\lambda \leq t_0$,

$$E_i(I) \leq E^J \Big(\big(\hat{y}_{J,i} - E^J \hat{y}_{J,i} \big)^2 \Big) \times \lambda^2 \bar{c}^2 \exp\Big(2\bar{c}\lambda(\tau + 1) \Big) \times \Big((\tau + 1)^2 H_2(2\bar{c}\lambda) + H_1(2\bar{c}\lambda) \Big).$$

Take λ small enough, say, $0 < \lambda \leq \lambda_0$, so that

$$\lambda^2 \bar{c}^2 \exp\Big(2\bar{c}\lambda(\tau+1)\Big)\Big((\tau+1)^2 H_2(2\bar{c}\lambda) + H_1(2\bar{c}\lambda)\Big) \leq \lambda \underline{c}/2$$

for $2\bar{c}\lambda \leq t_0$.

Thus, we have,

$$\begin{aligned}
&E_i\Big[\log E^J \exp\big(-\lambda L_{210}(y_i - \hat{y}_{J,i})\big)\Big] \\
&\leq -\lambda E_i L_{210}(y_i - \hat{y}_{j,i}) \\
&+ \lambda E_i\Big[L_{210}(y_i - \hat{y}_{j,i}) - E^J L_{210}(y_i - \hat{y}_{j,i})\Big] \\
&+ \lambda/2 E_i\Big[E^J L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \hat{y}_{j,i})\Big] \\
&\leq -\lambda E_i L_{210}(y_i - \hat{y}_{j,i}).
\end{aligned}$$

Further, similarly in Yang (2004),

$$\begin{aligned}
&-\lambda E \sum_{i=n_0}^{n_0+n-1} L_{210}(y_i - \hat{y}_j^*) \geq -E\log(1/q^n) \\
&\geq \log(1/\pi_j) - \lambda \sum_{i=1}^{n} E L_{210}(y_i - \hat{y}_{j,i}).
\end{aligned}$$

Since the analysis is based on an arbitrary j, so

$$\sum_{i=n_0}^{n_0+n-1} E L_{210}(y_i - \hat{y}_i^*) \leq \inf_{j\geq 1}\Big(\frac{\log(N)}{\lambda} + \sum_{i=n_0}^{n_0+n-1} E L_{210}(y_i - \hat{y}_{j,i})\Big).$$

This completes the proof of Theorem 1.

2.7.2 Proof of Theorem 2

For $\delta > 0$, recall

$$h(\delta) := \int \exp\Big(-\frac{L_{210}(x)}{\delta}\Big)dx.$$

Since

$$L_{210}(x) \le |x| + \frac{\alpha_1}{m}x^2 + \alpha_2 m, \quad L_{210}(x) \ge \frac{\alpha_1}{m}x^2, \tag{2.9}$$

then, from (2.9) and Condition 3,

$$\begin{aligned}
h(\delta) &\le \int \exp\Big(-\frac{\frac{\alpha_1}{m}x^2}{\delta}\Big)dx = \sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}},\\
h(\delta) &\ge \int \exp\Big(-\frac{|x| + \frac{\alpha_1}{m}x^2 + \alpha_2 m}{\delta}\Big)dx\\
&= \exp\Big(-\frac{m}{\delta}(\alpha_2 - \frac{1}{2\alpha_1})\Big)\int \exp\Big(-\frac{\alpha_1}{m\delta}(|x| + \frac{m}{2\alpha_1})^2\Big)dx\\
&= \sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\exp\Big(-\frac{m}{\delta}(\alpha_2 - \frac{1}{2\alpha_1})\Big)\\
&\times \int \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big(|\tilde{x}| + \sqrt{\frac{m}{2\alpha_1\delta}}\big)^2\Big)d\tilde{x}\\
&\ge \sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\exp\Big(\frac{m}{\delta}(\frac{1}{4\alpha_1} - \alpha_2)\Big)\xi_1,
\end{aligned}$$

where

$$0 < \xi_1 \le \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\big(-\frac{x^2}{2}\big)dx - \int_{-\frac{m}{2\alpha_1 A}}^{\frac{m}{2\alpha_1 A}}\frac{1}{\sqrt{2\pi}}\exp\big(-\frac{x^2}{2}\big)dx < 1.$$

Let $\xi_2 = \min(\exp\big(\frac{m}{A}\big(\frac{1}{4\alpha_1} - \alpha_2\big)\big)\xi_1, \exp\big(mA\big(\frac{1}{4\alpha_1} - \alpha_2\big)\big)\xi_1)$, then both ξ_1 and ξ_2 only depend on α_1, α_2, A and m. It follows that:

$$\sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\xi_2 \leq h(\delta) \leq \sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}. \tag{2.10}$$

Recall

$$g(x|\delta) := \frac{1}{h(\delta)} \exp\bigg(-\frac{L_{210}(x)}{\delta}\bigg). \tag{2.11}$$

Then, as in Yang (2004),

$$\sum_{i=1}^{n} ED(q_i||\hat{q}_i) = ED(f^n||q^n),$$

where

$$\begin{aligned}
f^n &= \prod_{i=1}^{n} \frac{1}{h(\delta_i)} \exp\bigg(-\frac{1}{\delta_i} L_{210}(y_i - \eta_i)\bigg) \\
&= \frac{1}{\prod_{i=1}^{n} h(\delta_i)} \exp\bigg(-\sum_{i=1}^{n} \frac{L_{210}(y_i - \eta_i)}{\delta_i}\bigg), \\
q^n &= \sum_{j=1}^{N} \frac{1}{N} \prod_{i=1}^{n} \frac{1}{h(\hat{\delta}_{j,i})} \exp\bigg(-\frac{1}{\hat{\delta}_{j,i}} L_{210}(y_i - \hat{y}_{j,i})\bigg) \\
&= \sum_{j=1}^{N} \frac{1}{N} \prod_{i=1}^{n} \frac{1}{h(\hat{\delta}_{j,i})} \exp\bigg(-\sum_{i=1}^{n} \frac{L_{210}(y_i - \hat{y}_{j,i})}{\hat{\delta}_{j,i}}\bigg).
\end{aligned}$$

Then

$$
\begin{aligned}
&\sum_{i=1}^{n} ED(q_i||\hat{q}_i) \\
&\leq E\log\left(\frac{\frac{1}{\prod_{i=1}^{n} h(\delta_i)}\exp\big(-\sum_{i=1}^{n}\frac{L_{210}(y_i-\eta_i)}{\delta_i}\big)}{\frac{1}{N}\prod_{i=1}^{n}\frac{1}{h(\hat{\delta}_{j,i})}\exp\big(-\sum_{i=1}^{n}\frac{L_{210}(y_i-\hat{y}_{j,i})}{\hat{\delta}_{j,i}}\big)}\right) \\
&= \log(N) + E\sum_{i=1}^{n}\left(\frac{L_{210}(y_i-\hat{y}_{j,i})}{\hat{\delta}_{j,i}} - \frac{L_{210}(y_i-\eta_i)}{\delta_i}\right) \\
&\quad + E\sum_{i=1}^{n}\log\left(\frac{h(\hat{\delta}_{j,i})}{h(\delta_i)}\right).
\end{aligned}
$$

From the Condition 3, there exists a positive constant $\xi_3 > 0$, such that:

$$
\left|\log\left(\frac{h(\hat{\delta}_{j,i})}{h(\delta_i)}\right)\right| \leq \xi_3|\hat{\delta}_{j,i}-\delta_i| \leq A\xi_3\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i} = c_1\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i} \tag{2.12}
$$

where $c_1 = A\xi_3$ and it depends on α_1, α_2, A and m.

Let E_i denotes the conditional expectation given z^{i-1}, it follows

$$
\begin{aligned}
&h(\delta_i)E_iL_{210}(y_i-\mu_i) \\
&= \int \exp\big(-\frac{1}{\delta_i}L_{210}(x)\big)L_{210}(x)dx \\
&\leq \int \exp\big(-\frac{\alpha_1}{m\delta_i}x^2\big)\frac{\alpha_1}{m}x^2dx + \int_{\frac{\alpha_1 A}{m}x^2\leq 1} L_{210}(x)dx \\
&\left(\text{For } \frac{\alpha_1 A}{m}x^2 \geq 1,\ \exp\big(-\frac{L_{210}(x)}{\delta_i}\big)L_{210}(x) \leq \exp\big(-\frac{\alpha_1 x^2}{m\delta_i}\big)\frac{\alpha_1 x^2}{m}\right) \\
&= 2\sqrt{\pi}\frac{m}{2\alpha_1}\delta_i^{3/2} + \xi_4\delta_i^{3/2} \\
&= \xi_5\delta_i^{3/2}
\end{aligned} \tag{2.13}
$$

where $\xi_4/A^{3/2} \geq \int_{\frac{\alpha_1 A}{m}x^2\leq 1} L_{210}(x)dx$ and $\xi_5 = \xi_4 + 2\sqrt{\pi}\frac{m}{2\alpha_1}$.

Then,

$$E_i L_{210}(y_i - \mu_i) \leq \frac{\xi_5 \delta_i^{3/2}}{\sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\xi_2} = \xi_6 \delta_i, \tag{2.14}$$

where $\xi_6 = \frac{\xi_5}{\sqrt{\frac{m\pi}{\alpha_1}}\xi_2}$ and it depends on α_1, α_2, A and m.
Further, from Condition 3, it follows:

$$\begin{aligned}
&\left|(\frac{1}{\hat{\delta}_{j,i}} - \frac{1}{\delta_i})L_{210}(y_i - \mu_i)\right| \leq \left|\frac{1}{\hat{\delta}_{j,i}} - \frac{1}{\delta_i}\right|\xi_6\delta_i \\
&\leq A^2\xi_6\frac{|\hat{\delta}_{j,i} - \delta_i|}{\delta_i} = c_2\frac{|\hat{\delta}_{j,i} - \delta_i|}{\delta_i},
\end{aligned} \tag{2.15}$$

where $c_2 = A^2\xi_6$ depending on α_1, α_2, A and m.

Similarly,

$$\left|\frac{1}{\hat{\delta}_{j,i}}\Big(L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)\Big)\right| \leq B\frac{|L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)|}{\delta_i}. \tag{2.16}$$

Therefore, from (2.12)-(2.16), it is true for any j that, for more details

$$\begin{aligned}
&\sum_{i=1}^{n} ED(q_i||\hat{q}_i) \\
&\leq \log(N) + \sum_{i=1}^{n}\Bigg(A^2 \times E\frac{|L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)|}{\delta_i} \\
&\quad + (c_1 + c_2)E\frac{|\hat{\delta}_{j,i} - \delta_i|}{\delta_i}\Bigg) \\
&\leq \log(N) + \sum_{i=1}^{n}\Bigg(A^3 \times E|L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)| \\
&\quad + A(c_1 + c_2)E|\hat{\delta}_{j,i} - \delta_i|\Bigg).
\end{aligned}$$

So,

$$\begin{aligned}
&\sum_{i=1}^{n} ED(q_i||\hat{q}_i) \\
&\leq \inf_j \Bigg(\log(N) + \sum_{i=1}^{n} \Bigg(CE|L_{210}(y_i - \hat{y}_{j,i}) - L_{210}(y_i - \eta_i)| \\
&\quad + CE|\hat{\delta}_{j,i} - \delta_i| \Bigg) \Bigg),
\end{aligned}$$

where $C \geq \max(A^3, A(c_1 + c_2))$ depends on α_1, α_2, A and m. This completes the proof of Theorem 2.

Chapter 3

Forecast Combination Under Heavy Tailed Errors

3.1 Introduction

When multiple forecasts are available for a target variable, well designed forecast combination methods can often outperform the best individual forecaster, as demonstrated in the literature of applications of forecast combinations in fields such as tourism, wind power generation, finance and economics in the last fifty years.

Many combination methods have been proposed from different perspectives since the seminal work of forecast combination by Bates & Granger (1969). See the discussions and summaries in Clemen (1989), Newbold & Harvey (2002) and Timmermann (2006) for key developments and many references. More recently, Lahiri et. al (2013) provided theoretical and numerical comparisons between adaptive and simple forecast combination methods. However, to our knowledge, few studies have proposed/discussed forecast combination methods that target at cases where the forecast errors exhibit heavy tail behaviors. In such situations, the familiar forecast combination methods such as simple average, least squares regression with or without constraints, or those based on variance-covariance of the forecasts, may perform very poorly (some numerical examples are provided in sections 3.4 and 3.5 in this

chapter). As a matter of fact, many important variables in finance, economics and other areas do have heavy tails. For example, Marinelli et. al (2001) discussed the evidences of heavy tailed distributions to model the exchange rates, and Harvey (2013) modeled the U.S. GDP with a Student's t distribution with a low degrees of freedom. Glasserman et. al (2002) wrote:

> *(Paragraph 4 of p. 240) Using different approaches to the problem and different sets of data, these studies consistently find high kurtosis and heavy tails. Moreover, most studies find that the tails in financial data are not so heavy as to produce infinite variance (as would be implied by a non normal stable distribution), though higher order moments (e.g., fifth and higher) may be infinite.*

Forecast combination methods are needed to handle the heavy tailed situations.

In this chapter, we propose two forecast combination methods following the spirit of the AFTER methods by Yang (2004). One is specially designed for situations when there is strong evidence that the forecast errors are heavy-tailed and can be modeled by a scaled Student's t-distribution. The other one is designed for more general uses. For the former case, we assume that the forecast errors follow a scaled Student's t-distribution with possibly unknown scaled parameter and degrees of freedom. For situations when the identification of the heaviness of tails of the forecast errors is not feasible, normal, double-exponential and scaled Student's t-distributions are considered at the same time as candidates for the distribution form of the forecast errors.

Technically, if the forecast errors are assumed to follow a normal or a double-exponential distribution with zero mean, then the conditional probability density functions used in the combining process of the AFTER scheme can be estimated relatively easily for all the candidate forecasters because the estimation of the conditional

scale parameters is straightforward. See, e.g., Zou & Yang (2004) and Wei & Yang (2012), for more details. However, this is not the same situation if a scaled Student's t-distribution is assumed. Among the literature discussing the maximum likelihood parameter estimation in Student's t-regressions in the last few decades, Fernandez & Steel (1999) and Fonseca et. al (2008) provided comprehensive summaries of the convergence properties of the parameter estimations in different situations. Both of them showed that the estimation of the degrees of freedom and the scale parameter simultaneously in a scaled Student's t-regression models suffers from monotonic likelihood because the likelihood goes to infinity as the scale parameter goes to zero if the degrees of freedom ν is not large enough. To deal with this difficulty, methods other than maximum likelihood estimation have been proposed in the literature. For example, one may fix the degrees of freedom first then estimate the scale parameter using method of moments or other tools (see, e.g., Kan & Zhou, 2003).

In this paper, we follow a two-step procedure to estimate the density function given a forecast error sequence: first, estimate the scale parameter for each element in a given candidate pool of degrees of freedom. Note that each combination of the degrees of freedom and the scale parameter leads to a different estimate of the density function. Second, the weight of a density estimate is assigned from its relative historical performance. The final density estimate is a weighted mean of all the candidate density estimates. More details about this procedure, including how to determine the pool of candidate estimates, are available in section 3.2. There are three major advantages of this procedure: first, because a pool of degrees of freedom (rather than a single candidate) is considered, it reduces the potential risk of picking a degrees of freedom parameter that is far from the truth. Second, the likelihood that each candidate density estimate is the best is purely decided by data. Third, the calculation of the combined estimator is easy and fast.

It is worth pointing out that some popular combination methods in the literature

make assumptions on the distributions of forecast errors that do not necessarily exclude heavy tailed behaviors. For example, methods that are based on the estimation of variance-covariance of forecasters require the existence of variances. Regression based forecast combination methods (see, e.g., Granger & Ramanathan, 1984) assume the existence of certain moments of the forecast errors. However, to our knowledge, these methods are not really designed to handle heavy-tailed errors and are not expected to work well for such situations.

Prior to our work, efforts have been made to deal with error distributions that have tails heavier than normal by adaptive forecast combination methods. For example, Sancetta (2010) assumed that the tails of the target variables are no heavier than exponential decays, which restrict the heaviness of the tails of the forecast errors. Wei & Yang (2012) designed a method for errors heavier than the normal distributions but not heavier than the double-exponential distributions. However, none of these methods can deal with forecast errors with tails as heavy as that of Student's t-distributions. The new AFTER methods in this paper will be shown to handle such situations.

The plan of the paper is as follows: section 3.2 introduces the forecast combination method designed for heavy-tailed error distributions; in section 3.3, a more general combination method is proposed. Simulations are presented in section 3.4, and section 3.5 provides a real data example. Section 3.6 includes a brief concluding discussion. The proofs of the theoretical results are in the appendix.

3.2 t-AFTER

In this section, we propose a forecast combination method when there is strong evidence that the innovation errors in the data-generating process are heavy-tailed and can be modeled by a scaled Student's t-distribution.

3.2.1 Problem Setting

Suppose at each time period $i \geq 1$, there are J forecasters available for predicting y_i and the forecast combination starts at $i_0 \geq 1$. Note that some combination methods may require i_0 to be large enough, e.g., 10, to give reasonably accurate combinations. Let $\hat{y}_{i,j}$ be the forecast of y_i from the j-th forecaster. Let $\hat{Y}_i := (\hat{y}_{i,1}, \cdots, \hat{y}_{i,J})$ be the vector of candidate forecasts for y_i made at time point $i-1$.

Suppose $y_i := m_i + \epsilon_i$, where m_i is the conditional mean of y_i given all available information prior to observing y_i and ϵ_i is the innovation error at time i. Assume ϵ_i is from a distribution with probability density function (*pdf*) $\frac{1}{s_i}h(\frac{x}{s_i})$, where s_i is the scale parameter that depends on the data before observing y_i and $h(\cdot)$ is a *pdf* with mean 0 and scale parameter 1.

Let $W_i := (W_{i,1}, \cdots, W_{i,J})$ be a vector of combination weights of $\hat{Y}_i$. It is assumed that $\sum_{j=1}^{J} W_{i,j} = 1$ and $W_{i,j} \geq 0$ for any $i \geq i_0$, $1 \leq j \leq J$. Let $W_{i_0} = (w_1, \cdots, w_J)$ be the initial weight vector. The combined forecast for y_i from a combination method is:

$$\hat{y}_i = \langle \hat{Y}_i, W_i \rangle, \tag{3.1}$$

where $\langle \mathtt{a}, \mathtt{b} \rangle$ stands for the inner-product of vectors $\mathtt{a}$ and $\mathtt{b}$. Specifically, when needed, we use a superscript δ on each W_i to denote the combination weights that correspond to the method δ. For example, in the following sections, $W_i^{A_2}$ and $W_i^{A_1}$ stand for the combination weights from the L_2- and L_1-AFTER methods, respectively.

3.2.2 The Existing AFTER Methods

As one recent method of adaptive forecast combination, the general scheme of adaptive forecast combination via exponential re-weighting (AFTER) was proposed by Yang (2004). It has been applied and studied in e.g., Fan et. al (2008), Inoue &

Kilian (2008), Sanchez (2008), Altavilla & Grauwe (2010), and Lahiri et. al (2013). Zhang et. al (2013) handled the case that the variable to be predicted is categorical.

In the general AFTER formulation, the relative cumulative predictive accuracies of the forecasters are used to decide their combining weights. Let $||\mathbf{x}||_1 := \sum_{i=1}^{n} |x_i|$ be the l_1-norm of vector $\mathbf{x} = (x_1, \cdots, x_n)$.

The general form of W_i for the AFTER approach is:

$$W_i = \frac{\mathbf{l}_{i-1}}{||\mathbf{l}_{i-1}||_1}, \tag{3.2}$$

where $\mathbf{l}_{i-1} = (l_{i-1,1}, \cdots, l_{i-1,J})$ and for any $1 \leq j \leq J$,

$$l_{i-1,j} = w_j \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j}} h\left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j}}\right), \tag{3.3}$$

where $\hat{s}_{i',j}$ is an estimate of $s_{i'}$ from the j-th forecaster at time point $i' - 1$.

To be more specific, the most commonly used AFTER procedures, the L_2-AFTER from Zou & Yang (2004) and the L_1-AFTER from Wei & Yang (2012), are given below with more details.

L_2-AFTER When the innovation errors in the data generating process follow a normal distribution (or a distribution close to a normal distribution), the L_2-AFTER is both theoretically and empirically competitive in providing combined forecasts that perform at least as well as any individual forecaster in any evaluation period plus a small penalty. Let f_N be the *pdf* of $N(0, 1)$. To get $W_i^{A_2}$, first use f_N as the h in expression (3.3), then plug the new $\mathbf{l}_{i-1}$ into (3.2). The $\hat{s}_{i,j}$ used in the L_2-AFTER, denoted as $\hat{\sigma}_{i,j}$, under iid assumption on the innovation errors, is the sample standard deviation of $\{y_{i'} - \hat{y}_{i',j}\}_{i'=1}^{i-1}$.

L_1-AFTER Let f_{DE} be the *pdf* of a double-exponential distribution with scale parameter 1 and location parameter 0. To get $W_i^{A_1}$ under the assumption that the

errors are from a double-exponential distribution, follow the same procedure for $W_i^{A_2}$ but use f_{DE} as the h in expression (3.3). The $\hat{s}_{i,j}$ used in the L_1-AFTER, denoted as $\hat{d}_{i,j}$, is the mean of $\{|y_{i'} - \hat{y}_{i',j}|\}_{i'=1}^{i-1}$.

3.2.3 The t-AFTER Methods

Since the estimation of the degrees of freedom and the scale parameter simultaneously in a scaled Student's t-regression setting suffers from certain theoretical difficulties as mentioned in the introduction, we use a different strategy in this paper. Specifically, we take an estimation procedure that has two steps: first, we estimate the scale parameter for each given degrees of freedom in a candidate pool; second, the relative weight of each estimated degrees of freedom and scale parameter pair is assigned from its relative historical performance. Let $\Omega := (\nu_1, \cdots, \nu_K)$ be a set of degrees of freedom for Student's t-distributions. The choice of Ω will be discussed later in this subsection. Let $w_{j,k}$ ($w_{j,k} \geq 0$ and $\sum_{k=1}^{K}\sum_{j=1}^{J} w_{j,k} = 1$) be the initial combination weight of the forecaster j under the degrees of freedom ν_k.

Let the combining weight of $\hat{Y}_i$ from a t-AFTER method be $W_i^{A_t}$ and let the combined forecast be $\hat{y}_i^{A_t}$. Then, $W_i^{A_t}$ and $\hat{y}_i^{A_t}$ are obtained via the following algorithm:

1. Estimate (e.g., by MLE) s_i for each $\nu_k \in \Omega$ and for each candidate forecaster. The estimate for s_i from the j-th forecaster given ν_k is denoted as $\hat{s}_{i,j,k}$.

2. Calculate $W_i^{A_t}$ and $\hat{y}_i^{A_t}$:

$$W_i^{A_t} = \frac{\mathbf{l}_{i-1}^{A_t}}{||\mathbf{l}_{i-1}^{A_t}||_1}, \quad \hat{y}_i^{A_t} = \langle \hat{Y}_i, W_i^{A_t} \rangle, \tag{3.4}$$

where $\mathbf{l}_{i-1}^{A_t} = (l_{i-1,1}^{A_t}, \cdots, l_{i-1,J}^{A_t})$ and for $1 \leq j \leq J$ and any $i \geq i_0 + 1$,

$$l_{i-1,j}^{A_t} = \sum_{k=1}^{K} l_{i-1,j,k}^{A_t} \quad \text{with} \quad l_{i-1,j,k}^{A_t} = w_{j,k} \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j,k}} f_t\left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j,k}} \middle| \nu_k\right), \tag{3.5}$$

where $f_t(\cdot|\nu)$ is the *pdf* of a Student's t-distribution with degrees of freedom ν.

It is assumed that the elements in Ω are natural numbers for the sake of convenience. In general, when no specific information is available to estimate the size of candidate degrees of freedom efficiently, one can start with a large but relatively sparse pool (say, $\{1, 3, 5, 8, 12, 15, 20, 30\}$) and then may narrow it down based on the performances on some training data sets. When there is strong evidence that the tails of the forecast errors are heavy, the size of Ω can be relatively small, say no more than 3 or 5. In this situation, from our experiences, $\Omega = \{1, 3\}$ or $\{1, 3, 5\}$ works well.

Obviously, when the innovation errors in the true model follow a scaled Student's t-distribution with a known degrees of freedom ν, then $\Omega := \{\nu\}$. Then expression (3.5) can be simplified into:

$$l_{i-1,j}^{A_t} = w_j \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j}} f_t \left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j}} \middle| \nu \right), \tag{3.6}$$

where w_j is the initial weight of the j-th forecaster and $\hat{s}_{i,j}$ is an estimate of s_i from the j-th forecaster using all information at and before time point $i-1$ when the true ν is known.

3.2.4 Risk Bounds of the t-AFTER

To avoid potential redundancy, we first give a risk bound on the t-AFTER assuming ν is known. A more general theorem that treats ν (and even the form of error distribution) as unknown will be given in section 3.3.

Conditions

Condition 4. There exists a constant $\tau > 0$ such that for any $i \geq i_0$,

$$\Pr(\sup_{1\leq j\leq J} |\hat{y}_{i,j} - m_i|/s_i \leq \sqrt{\tau}) = 1.$$

Condition 5. These exists a constant $\xi_1 > 0$ such that for any $i \geq i_0$ and $1 \leq j \leq J$:

$$\Pr\left(\frac{\hat{s}_{i,j}}{s_i} \geq \xi_1\right) = 1.$$

Condition 5′. These exists a constant $0 < \xi_1' < 1$ such that for any $i \geq i_0$ and $1 \leq j \leq J$:

$$\Pr\left(\xi_1' \leq \frac{\hat{s}_{i,j}}{s_i} \leq \frac{1}{\xi_1'}\right) = 1.$$

Condition 4 holds when the forecast errors are bounded, which is true in many real applications, although it excludes some time series models such as AR(1). It is required for the development of the theorems in this paper. See section 3.3.1 of Wei & Yang (2012) for more discussions on this condition.

Condition 5 generally requires that the estimates of the scale parameters are not too small compared to the truth. Condition 5′ requires that the estimates of the scale parameters are not too far from the truth in both directions.

Risk Bounds for the t-AFTER with a Known ν

Assume the true forecast errors follow a scaled Student's t-distribution with a known degrees of freedom ν. Let σ_i and s_i be the conditional standard deviation and scale parameter, respectively, of ϵ_i at time point i and let $\hat{s}_{i,j}$ be an estimator of s_i from the j-th forecaster.

Let $q_i = \frac{1}{s_i} f_t\left(\frac{y_i - m_i}{s_i}\middle|\nu\right)$ be the actual conditional error density function at time point i and $\hat{q}_i^{A_t} = \sum_{j=1}^{J} W_{i,j}^{A_t} \frac{1}{\hat{s}_{i,j}} f_t\left(\frac{\hat{y}_{i,j} - y_i}{\hat{s}_{i,j}}\middle|\nu\right)$, where $W_i^{A_t}$ is defined in expression

(3.4). So, $\hat{q}_i^{A_t}$ is the mixture estimator of q_i from the t-AFTER procedure. Let $D(f||g) := \int f \log \frac{f}{g}$ be the K-L divergence between two density functions f and g. So, $E\big(D(q_i||\hat{q}_i^{A_t})\big)$ is a measure of the performances of $\hat{q}_i^{A_t}$ as an estimate of q_i under the K-L loss at time point i.

Theorem 3

If the innovation errors are from a scaled Student's t-distribution with degrees of freedom ν and Condition 5 holds, then:

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} ED(q_i||\hat{q}_i^{A_t}) \leq \inf_{1\leq j\leq J}\left(\frac{\log\frac{1}{w_j}}{n} + \frac{1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_{i,j})^2}{2s_i^2} + \frac{B_1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(\hat{s}_{i,j}-s_i)^2}{s_i^2}\right).$$

Further, if ν is strictly larger than 2 and Conditions 4 and 5′ hold, then

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_i^{A_t})^2}{\sigma_i^2} \leq C\inf_{1\leq j\leq J}\left(\frac{\log\frac{1}{w_j}}{n} + \frac{B_2}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_{i,j})^2}{\sigma_i^2} + \frac{B_3}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(\hat{s}_{i,j}-s_i)^2}{s_i^2}\right).$$

In the above, C, B_1, B_2 and B_3 are constants. B_1 and B_3 depend on ξ_1 and ξ_1', respectively. B_2 is a function of ν and C depends on τ and ξ_1'. □

Remarks.

1. When only Condition 5 is satisfied, Theorem 3 shows that the cumulative distance between the true densities and their estimators from the t-AFTER is upper bounded by the cumulative (standardized) forecast errors of the best candidate forecaster plus a penalty that has two parts: squared relative estimation errors of the scale parameters and logarithm of the initial weights. This risk bound is obtained without assuming the existence of variances of the random errors and $\hat{s}_{i,j}/s_i$ is only required to be lower-bounded.

2. When ν is assumed to be strictly larger than 2 and both Conditions 4 and 5′ are satisfied, Theorem 3 shows that the cumulative forecast errors have the

same convergence rate of the cumulative forecast errors of the best candidate forecaster plus a penalty that depends on the initial weights and efficiency of scale parameters estimation. The risk bounds hold even if the the distribution of random errors have tails as heavy as t_3.

3. If there is no prior information to decide the w_j's in expression (3.6), then equal initial weights could be applied. That is, $w_j = 1/J$ for all j. In this case, it is easy to see that the number of candidate forecasters plays a role in the penalty. When the candidate pool is large, some preliminary analysis should be done to eliminate the significantly less competitive ones before applying the t-AFTER.

The proof of Theorem 3 is in sections 3.7.1-3.7.3.

3.3 g-AFTER

In section 3.2, the t-AFTER provides theoretically justified forecast combination when the random errors are known to have Student's t-distributions. However, the error distribution is typically unknown.

In this section, we propose a forecast combination method, g-AFTER, for situations when there is a lack of strong or consistent evidence on the tail behaviors of the forecast errors due to shortage of data and/or evolving data-generating process. A theorem that allows the random errors to be from one of the three popular distribution families (normal, double-exponential, and scaled Student's t) is provided to characterize the performance of the g-AFTER.

3.3.1 The g-AFTER Method

Let the combining weight of $\hat{Y}_i$ from the g-AFTER be $W_i^{A_g}$. For any $i > i_0$, $W_i^{A_g}$ and the associated combined forecast $\hat{y}_i^{A_g}$ are:

$$W_i^{A_g} = \frac{\mathbf{l}_{i-1}^{A_g}}{||\mathbf{l}_{i-1}^{A_g}||_1}, \quad \hat{y}_i^{A_g} = \langle \hat{Y}_i, W_i^{A_g} \rangle, \tag{3.7}$$

where $\mathbf{l}_{i-1}^{A_g} = (l_{i-1,1}^{A_g}, \cdots, l_{i-1,J}^{A_g})$ and for $1 \leq j \leq J$,

$$l_{i-1,j}^{A_g} = l_{i-1,j}^{A_2} + c_1 l_{i-1,j}^{A_1} + c_2 l_{i-1,j}^{A_t}, \tag{3.8}$$

where $l_{i-1,j}^{A_2}$, $l_{i-1,j}^{A_1}$ and $l_{i-1,j}^{A_t}$ are from the L_2-, L_1- and t-AFTERs, respectively and c_1 and c_2 are non-negative constants that control the relative importances of the L_2-, L_1- and t-AFTERs in the g-AFTER. For instance, c_1 and c_2 can be small when one has evidence that suggests the innovation errors are likely to be normally distributed.

3.3.2 Conditions

Condition 6. Suppose the random errors have zero mean and are from one of the three families (normal, double exponential, and scaled Student's t), and there exists a constant $0 < \xi_2 \leq 1$ such that for any $i \geq i_0$, with probability 1, we have

$$\xi_2 \leq \frac{\hat{s}_i}{s_i} \leq \frac{1}{\xi_2},$$

where s_i the actual conditional scale parameter at time point i and $\hat{s}_i$ refers to any estimate of s_i used in the g-AFTER.

This condition requires all the estimates of the scale parameters stay in a reasonable range around the true values. For the j-th candidate forecaster, $\hat{s}_i$ is $\hat{\sigma}_{i,j}$ when associated with normal errors, is $\hat{d}_{i,j}$ when associated with the double exponential,

and is $\hat{s}_{i,j,k}$ when associated with the scaled Student's t with degrees of freedom ν_k, where $\hat{\sigma}_{i,j}$, $\hat{d}_{i,j}$, $\hat{s}_{i,j,k}$ and ν_k are defined in section 3.2.2 and 3.2.3.

Condition 7. When the innovation errors in the true model follow a scaled Student's t-distribution with degrees of freedom ν, assume there exist positive constants $\underline{\nu}$, λ and $\bar{\nu}$ such that,

$$\underline{\nu} \leq \min_{\nu_k \in \Omega}(\nu_k, \nu) - 2 \leq \bar{\nu}, \quad \max_{\nu_k \in \Omega} |\nu_k - \nu| \leq \lambda.$$

3.3.3 Risk Bounds for the g-AFTER

Let $w_j^{A_2}$ and $w_j^{A_1}$ be the initial combination weights of the forecaster j in the L_2- and L_1-AFTERs respectively and $w_{j,k}^{A_t}$ be the initial combination weight of the j-th forecaster under the degrees of freedom ν_k in the t-AFTER.

Let $\hat{W}_{i,j}^{A_2} = \frac{l_{i-1,j}^{A_2}}{||\mathbf{l}_{i-1}^{A_g}||_1}$, $\hat{W}_{i,j}^{A_1} = \frac{c_1 l_{i-1,j}^{A_1}}{||\mathbf{l}_{i-1}^{A_g}||_1}$ and $\hat{W}_{i,j,k}^{A_t} = \frac{c_2 l_{i-1,j,k}^{A_t}}{||\mathbf{l}_{i-1}^{A_g}||_1}$, where $l_{i-1,j,k}^{A_t}$ is defined in expression (3.5) and $\mathbf{l}_{i-1}^{A_g}$ is defined in expression (3.8). So, $\hat{W}_{i,j}^{A_2}$, $\hat{W}_{i,j}^{A_1}$ and $\hat{W}_{i,j,k}^{A_t}$ are the weights of the density estimates under normal, double-exponential and scaled Student's t with degrees of freedom ν_k in the g-AFTER procedure at time point $i-1$ from the j-th forecast, respectively. Let $G = \sum_{j=1}^{J}(w_j^{A_2} + c_1 w_j^{A_1} + c_2 \sum_k w_{j,k}^{A_t})$, where c_1 and c_2 are defined in expression (3.8).

Let q_i be the *pdf* of ϵ_i at time point i and its estimator from a g-AFTER procedure be:

$$\hat{q}_i^{A_g} = \sum_{j=1}^{J}\left(\hat{W}_{i,j}^{A_2}\frac{1}{\hat{\sigma}_{i,j}}f_N\left(\frac{\hat{y}_{i,j}-y_i}{\hat{\sigma}_{i,j}}\right) + \hat{W}_{i,j}^{A_1}\frac{1}{\hat{d}_{i,j}}f_{DE}\left(\frac{\hat{y}_{i,j}-y_i}{\hat{d}_{i,j}}\right) + \sum_{k=1}^{K}\hat{W}_{i,j,k}^{A_t}\frac{1}{\hat{s}_{i,j,k}}f_t\left(\frac{\hat{y}_{i,j}-y_i}{\hat{s}_{i,j,k}}|\nu_k\right)\right).$$

Theorem 4

If Conditions 6 and 7 hold, then for $\hat{y}_i^{A_g}$ from a g-AFTER procedure, we have:

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} ED(q_i||\hat{q}_i^{A_g}) \le \inf_{1\le j\le J}\left(\frac{B_1}{n}\sum_{i=i_0+1}^{i_0+n} E\left(\frac{(m_i-\hat{y}_{i,j})^2}{\sigma_i^2}\right)+R\right),$$

where

$$R=\begin{cases}\frac{\log\left(\frac{G}{w_j^{A_2}}\right)}{n}+\frac{B_2}{n}\sum_{i=i_0+1}^{i_0+n}E\frac{(\hat{\sigma}_{i,j}-\sigma_i)^2}{\sigma_i^2}, & \text{under normal errors;}\\ \frac{\log\left(\frac{G}{c_1 w_j^{A_1}}\right)}{n}+\frac{B_2}{n}\sum_{i=i_0+1}^{i_0+n}E\frac{(\hat{d}_{i,j}-d_i)^2}{d_i^2}, & \text{under double-exponential err}\\ \inf_{1\le k\le K}\left(\frac{\log\left(\frac{G}{c_2 w_{j,k}^{A_t}}\right)}{n}+\frac{B_2}{n}\sum_{i=i_0+1}^{i_0+n}E\frac{(\hat{s}_{i,j,k}-s_i)^2}{s_i^2}+B_3\left|\frac{\nu-\nu_k}{\nu}\right|\right), & \text{under scaled } t \text{ errors.}\end{cases}$$

If Condition 4 also holds, then

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_i^{A_g})^2}{\sigma_i^2} \le C\inf_{1\le j\le J}\left(\frac{B_1}{n}\sum_{i=i_0+1}^{i_0+n}E\left(\frac{(m_i-\hat{y}_{i,j})^2}{\sigma_i^2}\right)+R\right).$$

In the above, C, B_1, B_2 and B_3 are constants depending on τ, ξ_2 and parameters in Condition 7. □

Remarks.

1. Theorem 4 provides a risk bound for more general situations compared to Theorem 3. That is, as long as the the true random errors are from one of the three popular families, similar risk bounds hold.

2. When strong evidence is shown that the errors are highly heavy-tailed, Ω can be very small with only small degrees of freedom and the $c_2 w_{j,k}^{A_t}$ in G can be

relatively large (relative to $w_j^{A_2}$ and $c_1 w_j^{A_1}$). The more information on the tails of the error distributions is available, the more efficient the allocation of the initial weights can be.

3. Specially, when the true random errors have tails significantly heavier than normal and double-exponential, they could be assumed to be from a scaled Student's t-distribution with unknown ν and a (general) t-AFTER procedure is more reasonable. In this case, $l_{i-1,j}^{A_g} = l_{i-1,j}^{A_t}$.

   Let $q_i = \frac{1}{s_i} f_t\left(\frac{\hat{y}_{i,j}-y_i}{s_i}\right)$ and $\hat{q}_i^{A_t} = \sum_{j,k} \hat{w}_{i,j,k}^{A_t} \frac{1}{\hat{s}_{i,j,k}} f_t\left(\frac{\hat{y}_{i,j}-y_i}{\hat{s}_{i,j,k}}\middle|\nu_k\right)$ and $\hat{w}_{i,j,k}^{A_t} \geq 0$ for all j and k. Without assuming Condition 4 is satisfied, it follows for any $n \geq 1$:

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} ED(q_i||\hat{q}_i^{A_t}) \leq \inf_{1\leq j\leq J}\left(\frac{\log(1/w_{i,j}^{A_t})}{n} + \frac{B_1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_{i,j})^2}{\sigma_i^2} + R^*\right),$$

   where $w_{j,k}^{A_t}$ is defined the same as that in section 3.2.3 and

$$R^* = \inf_{1\leq k\leq K}\left(\frac{B_2}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(\hat{s}_{i,j,k}-s_i)^2}{s_i^2} + B_3\left|\frac{\nu-\nu_k}{\nu}\right|\right).$$

   If Condition 4 is also satisfied, then it follows:

$$\frac{1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_i^{A_t})^2}{\sigma_i^2} \leq C\inf_{1\leq j\leq J}\left(\frac{\log(1/w_{i,j}^{A_t})}{n} + \frac{B_1}{n}\sum_{i=i_0+1}^{i_0+n} E\frac{(m_i-\hat{y}_{i,j})^2}{\sigma_i^2} + R^*\right),$$

   where C, B_1, B_2 and B_3 are the same as in Theorem 4.

The proof of Theorem 4 is in section 3.7.4.

3.4 Simulations

We consider two simulation scenarios, with candidate forecasters from linear regression models and autoregressive (AR) models. Results from the linear regression models show improvements of the t- and g-AFTERs over the L_1- and L_2-AFTERs when the innovation errors have heavy tails. In the AR settings, the t- and g-AFTERs are compared to many other popular combination methods in various situations, including cases that the forecast errors are with extremely symmetric/asymmetric heavy tails. We also compared the performances of the t- and g-AFTERs to other combination methods on the linear regression models and similar results are found. Only representative results are given here.

In this and the following sections, we have the following settings:

- Use $\Omega = \{1, 3\}$. The t-AFTER is proposed mostly to be applied when the error terms exhibit very strong heavy-tailed behaviors. When the degrees of freedom of the Student's t-distribution gets larger, the t-AFTER becomes similar to the L_1- or L_2-AFTER. Thus a choice of Ω with relatively small degrees of freedom in the g-AFTER should provide good enough adaption capability. In fact, other options for Ω, such as $\Omega = \{1, 3, 5, 8, 15\}$ were considered, and similar results were found.

- Since it is usually the case that g-AFTER is preferred when the users have no consistent and strong evidences to identify the distribution of the error terms from the three candidate distribution families, we put equal initial weights to the candidate distributions. So $c_1 = 1$, $c_2 = 2$, $w_j^{A_1} = w_j^{A_2} = 1/J$ and $w_{j,k}^{A_t} = \frac{1}{2J}$ are used in the g-AFTER. Note that, for example, if there is clear and consistent evidence that the error distribution is more likely to be from the normal distribution family, then putting relatively large initial weights on the L_2-AFTER procedure in a g-AFTER can be more appropriate than using equal

weights.

- The $\hat{s}_{i,j,k}$'s are the sample median of the absolute forecast errors before time point i from the forecaster j divided by the theoretical median of the absolute value of a random variable with distribution t_{ν_k}.

3.4.1 Linear Regression Models

Simulation Settings

There are p predictors $(X_1, \cdots, X_p)$ available and the true model uses the first p_0 predictors with coefficients $\beta = (\beta_1, \cdots, \beta_{p_0})$. That is, $Y = \sum_{i=1}^{p_0} X_i\beta_i + \epsilon$. The p candidate forecasters are generated from the following p models: $Y = \beta_0 + X_1\beta_1 + e$, $Y = \beta_0 + \sum_{i=1}^{2} X_i\beta_i + e$, $\cdots$, $Y = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + e$. We take $p = 2p_0 - 1$ for this scenario. Other settings for p and p_0 were also considered and they gave similar results.

The p predictors are generated from a multivariate normal distribution with zero mean and covariance matrix Σ with sample size $n = 125$. For the entries in Σ, the diagonal elements are 1 and off-diagonal elements are 0.8. The forecasters are generated after the 90-th observation, and the combination is generated after the 5th forecasts. Various distributions for the random errors (ϵ) are considered. Note that, we also tried other structures of Σ, including the ones with $\Sigma_{i,j} = 0.5^{|i-j|}$ and $\Sigma_{i,j} = I(i = j)\ \forall 1 \leq i, j \leq p$. The results are similar.

For each set of β, we generate 200 sets of $(X_1, \cdots, X_p, Y)$ and on each of the 200 sets, we record the $\frac{1}{20}\sum_{i=106}^{125}(m_i - \hat{y}_i)^2$ (Average Squared Estimation Error (ASEE hereafter)) of each combination method, where $\hat{y}_i$ is the forecast of y_i from this method. Note that, since this is a simulation study, the combined forecasts are compared with the conditional means (m_i's) instead of the observations (y_i's) to better compare the competing methods. For each competing method, the mean ASEE

over the 200 data sets is recorded.

We sample β for 200 times independently from a $Unif[1,3]$ for each component with size p_0, so 200 sets of mean ASEEs are recorded. In order to compare the performances of the four AFTER based methods, the L_2-, L_1-, t- and g-AFTERs, for each β, the ratios of the mean ASEEs of the L_2-, t- and g-AFTERs over the mean ASEE of the L_1-AFTER is recorded. The summaries (means and their standard errors) of the 200 sets of ratios are presented.

Results

Three sets of results ($p_0 = 3, 5, 10$ respectively) are presented in Table 3.1 in this subsection. In this table, `A2`, `At` and `Ag` stand for the ratios of the mean ASEEs of the L_2-, t- and g-AFTERs over those of the L_1-AFTER. The information in the first and second rows indicate the distributions of ϵ: t_3 with $\sigma^2 = 9$ means $\epsilon \sim kt_3$ with $Var(kt_3) = 9$. The top numbers in rows 4-6, 8-10 and 12-14 are the mean of the 200 ratios. The numbers in the parentheses are the standard errors of the statistics above them. Rows 3, 7 and 11 tell the number of predictors used in the true models. DE stands for double-exponential with zero mean hereafter.

Summary

From Table 3.1, in the linear regression setting, we see that the overall performances of the t- and g-AFTERs are relatively more robust than that of the L_1- and L_2-AFTERs. Specifically:

1. When the random errors have heavy tails, the t- and g-AFTERs provide more accurate forecasts than the L_2- and L_1-AFTERs consistently.

2. When the tails of the random errors distributions are not or only mildly heavy, say a normal or a scaled Student's t-distribution with a large degrees of freedom, the g-AFTER is better than the t-AFTER in terms of forecast accuracy.

3. The L_1-AFTER outperforms the L_2-AFTER when the random errors have heavy tails while L_2-AFTER is more accurate than the L_1-AFTER when the random errors are not heavy-tailed.

3.4.2 AR Models

Simulation Settings

Let the true model be a $AR(p_0)$ process with innovation errors from certain distributions and the candidate forecasters be based on $AR(1), AR(2), \cdots, AR(p)$ $(1 \leq p_0 \leq p)$, respectively. For results on asymptotically optimal model selection for AR models, see, e.g., Ing (2007) and Ing et. al (2012). We here compare forecast combination methods.

In this scenario, given p, p_0 is randomly sampled from a Uniform distribution on $\{1, 2, \cdots, p\}$. Given p_0, β in the true model is generated; given β, 200 samples with size $n = 125$ from the true model are generated. On each data sample, the candidate forecasters are generated after the 90-th observation and the ASEE of the last 20 forecasts is recorded. Also, the combined forecasts are compared with the conditional

Table 3.1: Simulation Results on the Linear Regression Models

	t_3		DE		t_{10}		$normal$	
	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 9$
	$p_0 = 3$							
$A2$	1.302	1.043	1.116	1.028	0.983	0.958	0.926	0.931
	(0.009)	(0.003)	(0.004)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)
At	0.943	0.980	0.983	0.995	0.941	0.955	0.932	0.942
	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
Ag	0.944	0.967	0.974	0.977	0.940	0.950	0.926	0.938
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	$p_0 = 5$							
$A2$	1.257	1.066	1.088	1.026	0.980	0.955	0.937	0.927
	(0.008)	(0.004)	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
At	0.950	0.967	0.976	0.982	0.951	0.950	0.943	0.938
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Ag	0.951	0.958	0.971	0.970	0.949	0.944	0.939	0.933
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	$p_0 = 10$							
$A2$	1.166	1.056	1.035	0.998	0.968	0.949	0.946	0.929
	(0.006)	(0.003)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
At	0.950	0.957	0.964	0.965	0.949	0.946	0.948	0.939
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Ag	0.945	0.949	0.961	0.955	0.944	0.939	0.942	0.933
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

means instead of the observations. For each β, the mean ASEE of each combining method over the 200 samples is recorded and ratios of the mean ASEEs of other methods over that of the L_1-AFTER are recorded.

We replicate the generation of p_0's (and β's) for 200 times and report the mean and its standard error of the 200 ratios for each combination method.

Only the results of $p = 5$ are presented (other choices, such as $p = 8$ and 10, provide similar results).

Other Combination Methods

Some other popular combination methods are included in this part and compared with the newly proposed methods. Simple average combination strategy (`SA`) uses the average of the candidate forecasts as the combined forecasts. The `MD` and `TM` strategies use the median and the trimmed mean (remove the largest and smallest before averaging) of candidate forecasts, respectively. The variance-covariance estimation based combination method (denoted as `BG` because it was first proposed by Bates & Granger (1969)) we use in this paper is the version in Hansen (2008). Also, a modified `BG` method with a discount factor $0 < \rho < 1$ is considered and the results of multiple ρ's are presented. In the modified `BG`, the estimate of the (conditional) variance of the forecast errors of a forecaster at any time point is the associated discounted mean squared forecast error with factor ρ. See, e.g, Stock & Watson (2006), for more details. Hereafter, for example, $\texttt{BG}_{0.9}$ denotes a `BG` method with $\rho = 0.9$. Two linear-regression based combination methods are also considered: one is the combination via ordinary linear regression (`LR`) and the other one is a constrained linear regression (`CLR`) combination. The constraints of the `CLR` are: all coefficients are non-negative and the sum of the coefficients is 1 (without intercept in the regressions).

Results

Table 3.2 and Table 3.3 provide the summaries of the simulation results. In these two tables, `A2`, `At`, `Ag`, `SA`, `MD`, `TM`, `BG`, `LR` and `CLR` stand for the relative performances of these methods over that of the L_1-AFTER. The other entries are defined as in Table 3.1.

Table 3.2 presents the results for the cases that the innovation errors are not (or only mildly) heavy-tailed, while Table 3.3 contains the results when the random errors have significant heavy tails.

Table 3.2: Simulation Results on the AR Models with $p = 5$ (not or only mildly heavy tailed)

	normal			t_{10}			DE		
	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 9$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 9$
$A2$	0.941	0.940	0.940	0.972	0.972	0.971	1.030	1.032	1.033
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
At	0.954	0.953	0.954	0.961	0.962	0.962	0.997	1.001	0.995
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Ag	0.948	0.947	0.948	0.957	0.959	0.958	0.978	0.983	0.976
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
SA	2.892	2.484	2.408	2.372	2.297	2.070	2.278	2.176	2.483
	(0.268)	(0.166)	(0.189)	(0.167)	(0.174)	(0.127)	(0.148)	(0.151)	(0.148)
MD	1.681	2.025	1.824	1.884	1.874	1.421	1.740	1.602	1.943
	(0.137)	(0.191)	(0.187)	(0.243)	(0.197)	(0.076)	(0.137)	(0.144)	(0.168)
TM	1.805	1.946	1.754	1.838	1.705	1.469	1.723	1.571	1.885
	(0.121)	(0.144)	(0.134)	(0.156)	(0.138)	(0.066)	(0.109)	(0.093)	(0.120)
BG	1.441	1.462	1.389	1.425	1.364	1.321	1.431	1.357	1.500
	(0.047)	(0.051)	(0.047)	(0.042)	(0.040)	(0.032)	(0.046)	(0.035)	(0.045)
$BG_{0.95}$	1.432	1.453	1.381	1.417	1.358	1.315	1.427	1.353	1.495
	(0.047)	(0.050)	(0.047)	(0.042)	(0.040)	(0.032)	(0.045)	(0.035)	(0.045)
$BG_{0.9}$	1.429	1.449	1.378	1.414	1.355	1.313	1.425	1.352	1.492
	(0.047)	(0.049)	(0.047)	(0.042)	(0.039)	(0.032)	(0.045)	(0.035)	(0.045)
$BG_{0.8}$	1.433	1.452	1.382	1.417	1.357	1.315	1.427	1.353	1.491
	(0.047)	(0.050)	(0.047)	(0.042)	(0.040)	(0.032)	(0.045)	(0.035)	(0.044)
$BG_{0.7}$	1.447	1.464	1.394	1.428	1.366	1.322	1.432	1.357	1.495
	(0.048)	(0.051)	(0.049)	(0.043)	(0.040)	(0.033)	(0.046)	(0.036)	(0.045)
LR	7.956	8.355	8.491	8.856	10.210	9.138	11.110	11.240	10.040
	(0.346)	(0.339)	(0.342)	(0.387)	(1.032)	(0.363)	(0.504)	(0.509)	(0.513)
CLR	1.036	1.024	1.036	1.032	1.036	1.042	1.072	1.070	1.045
	(0.011)	(0.013)	(0.012)	(0.011)	(0.010)	(0.011)	(0.011)	(0.011)	(0.013)

Summary

In the autoregression scenario, we see that the t- and g-AFTERs consistently outperform all other non-AFTER based combination methods in all the simulated situations (heavy tailed or not) and outperform the L_1- and L_2-AFTERs when the innovation errors are not normal. Below are some important details:

1. In between the t- and g-AFTER, the latter is more robust since its performances under all scenarios are the best or close to the best. For the t-AFTER, its advantages over the L_1- and L_2-AFTERs are clear when the tails of the distributions of the innovation errors get heavier.

2. In both Table 3.2 and Table 3.3, the `CLR` is the most competitive method outside the AFTER family. It is because the constraints in the `CLR` make its weights relatively more stable and resistant to dramatic changes. The `CLR` gets more competitive when the innovation errors have heavier tails.

3. The `SA` and `TM` are vulnerable to outliers, which hurts their overall performances. We can see this from both tables.

4. In our settings, similar to many real application situations, since some of the candidate forecasters are highly correlated, using only the conditional variances to assign relative combining weights may not be enough. This explains why the `BG` and the discounted `BG`'s are not quite competitive as seen in Table 3.2 and Table 3.3.

Table 3.3: Simulation Results on the AR Models with $p = 5$ (heavy tailed)

	t_3			log-normal		
	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 9$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1$
$A2$	1.058	1.056	1.053	0.964	1.024	1.051
	(0.009)	(0.008)	(0.008)	(0.003)	(0.004)	(0.010)
At	0.955	0.947	0.961	0.951	0.940	0.921
	(0.006)	(0.006)	(0.006)	(0.003)	(0.004)	(0.008)
Ag	0.950	0.943	0.957	0.950	0.946	0.926
	(0.006)	(0.006)	(0.006)	(0.003)	(0.004)	(0.008)
SA	2.047	1.889	1.931	2.253	2.143	1.730
	(0.107)	(0.098)	(0.139)	(0.173)	(0.115)	(0.087)
MD	1.692	1.396	1.657	1.517	1.441	1.370
	(0.135)	(0.066)	(0.182)	(0.097)	(0.085)	(0.078)
TM	1.625	1.438	1.508	1.559	1.555	1.404
	(0.091)	(0.060)	(0.112)	(0.086)	(0.080)	(0.057)
BG	1.369	1.307	1.286	1.329	1.374	1.278
	(0.034)	(0.025)	(0.033)	(0.039)	(0.038)	(0.025)
$BG_{0.95}$	1.365	1.303	1.282	1.322	1.370	1.275
	(0.033)	(0.025)	(0.033)	(0.038)	(0.038)	(0.025)
$BG_{0.9}$	1.360	1.299	1.277	1.319	1.367	1.271
	(0.033)	(0.025)	(0.032)	(0.037)	(0.037)	(0.024)
$BG_{0.8}$	1.352	1.290	1.269	1.320	1.366	1.259
	(0.032)	(0.024)	(0.030)	(0.038)	(0.037)	(0.023)
$BG_{0.7}$	1.345	1.284	1.263	1.327	1.368	1.248
	(0.032)	(0.023)	(0.030)	(0.039)	(0.037)	(0.023)
LR	95.280	38.290	46.220	9.316	13.180	174.000
	(60.670)	(7.566)	(9.192)	(0.375)	(0.891)	(56.286)
CLR	1.014	1.007	1.016	1.046	1.032	0.974
	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.010)

`Note`: For the columns of 'log-normal', σ's are the scale parameters.

3.5 Real Data Example

In this section, we show advantages of the newly proposed methods over many popular competing combination methods on real data. Based on the data, two major comparisons are conducted: one compares the methods on a large collection of variables and the other on a focused subset. Since the t- and g-AFTERs are proposed to be the alternatives to the L_1-AFTER when heavy-tailed errors are present, then it is of particular interests to see their performances when L_1-AFTER fails to perform well.

3.5.1 Data and Settings

The M3-competition data are very popular in the field of predictive modeling (see, e.g., Makridakis & Hibon, 2000) and forecast combination. It contains 3003 financial/economical variables in which 1428 (N1402-N2829) have 18 forecasts and the rest have only 6 or 8 forecasts. For each of the 3003 variables, notice that the forecasts are generated all at once (1-, 2-,$\cdots$ and up to 6, 8 or 18-step ahead) by each forecaster. There were 24 candidate forecasters for each of the variables. We use the 1428 variables with 18 forecasts because some combination methods (such as the BG, $A2$ and so on) need a few forecasts to train the parameters before achieving a reasonable level of reliability.

Let $\hat{y}_{i'}$ be the forecast of $y_{i'}$ for $n_0 \leq i' \leq n_1$, then the mean squared forecast error (MSFE) is $\frac{1}{n_1-n_0+1}\sum_{i=n_0}^{n_1}(y_i - \hat{y}_i)^2$. Since the true model is not available for any of these 1428 variables, we use the mean squared forecast errors to measure the prediction performances.

Also, a more specific subset of N1402-N2829 is considered. On each variable in this subset that contains 23 variables, the MSFE of the L_1-AFTER is at least twice that of the `SA`.

Specifically, using the same notations as those in section 3.4.2, the averaged relative performances (MSFE) of the `MD`, `TM`, `BG`, discounted `BG`'s, `A2`, `A1`, `At` and `Ag` over the `SA` over the 1428 variables are presented. The main reason that we use the `SA` as the benchmark on this real data set is that the `SA` is one of the most popular combination methods with a great reputation in a broad range of applications. Since there are too many candidate forecasters compared to the forecast periods available, the two linear regression related combination methods discussed in section 3.4.2 are not considered here.

For each of the variables with 18 forecast periods, the combination starts after the 6-th forecasts and the MSFE of the last 9 forecasts of each method is recorded for performance comparisons. For each variable, the MSFE ratio of each method over that of the `SA` is reported. The summaries, mean (and its standard error), median, minimum, the $1st$, $3rd$ quartiles (denoted as Q_1 and Q_3, respectively) and maximum, of the 1428 ratios of each method are reported in Table 3.4 and Table 3.5.

Table 3.4: Results on the 1428 Variables of the M3-Competition Data

	mean	se	median	min	Q_1	Q_3	max
MD	1.050	0.010	1.022	0.002	0.910	1.143	5.341
TM	0.990	0.004	1.000	0.002	0.974	1.023	2.437
BG	0.784	0.010	0.838	0.001	0.596	0.973	5.227
$BG_{0.95}$	0.775	0.010	0.832	0.001	0.582	0.969	7.715
$BG_{0.9}$	0.768	0.012	0.825	0.001	0.564	0.966	11.45
$BG_{0.8}$	0.758	0.019	0.806	0.001	0.529	0.960	24.08
$BG_{0.7}$	0.757	0.031	0.793	0.001	0.503	0.956	43.19
$A1$	0.708	0.016	0.649	0.001	0.307	0.994	11.50
$A2$	0.697	0.017	0.639	0.001	0.309	0.979	13.32
At	0.708	0.015	0.646	0.001	0.312	1.003	8.632
Ag	0.696	0.014	0.645	0.001	0.308	0.987	7.710

Table 3.5: Results on the 23 Variables of the M3-Comptition Data

	mean	se	median	min	Q_1	Q_3	max
MD	1.822	0.181	1.491	0.666	1.154	2.310	4.038
TM	1.141	0.088	1.012	0.702	0.919	1.069	2.437
BG	1.990	0.197	1.835	0.632	1.283	2.379	5.110
$BG_{0.95}$	0.172	0.826	1.843	0.595	1.277	2.359	4.523
$BG_{0.9}$	0.152	0.731	1.802	0.562	1.284	2.154	4.027
$BG_{0.8}$	0.128	0.613	1.682	0.512	1.322	1.951	3.294
$BG_{0.7}$	0.115	0.551	1.494	0.482	1.195	1.842	2.836
$A1$	3.409	0.441	2.693	2.054	2.201	3.437	11.504
$A2$	3.448	0.569	2.387	1.526	2.039	3.209	13.317
At	3.145	0.389	2.562	1.058	2.179	3.205	8.632
Ag	2.971	0.329	2.379	1.059	2.136	3.137	7.710

3.5.2 Summary

General relative performances of `MD`, `TM`, `BG`, discounted `BG`'s, `A2`, `A1`, `At` and `Ag` to `SA` on N1402-N2829 are presented in Table 3.4. The results on the 23 variables are summarized in Table 3.5.

1. From Table 3.4, the overall performances of the AFTER based methods are better than the other popular combination methods considered. It also shows that the AFTERs can occasionally be significantly worse than the `SA` and other methods.

2. From Table 3.4, it is worth noticing that the performances of the AFTERs can be a thousand times better while only about 10 times worse than that of `SA`. An examination reveals that for certain variables, such as N1837 and N2217, some candidate forecasters are consistently and significantly worse than others. In this situation, since the `SA` can not remove the extreme 'disturbing' ones before averaging, its performance is extremely poor. However, the AFTERs essentially ignore the 'unreasonable' candidate forecasts so they can be significantly better than the `SA`.

3. Table 3.4 suggests that the t- and g-AFTERs have competitive performances in general while being more robust than others since their overall performances are outstanding and are still acceptable for the worst cases.

4. From the comparison in Table 3.5, the improvements over the L_1- and L_2-AFTERs of the t- and g-AFTERs are reasonably significant, although they all are still not as good as the `SA`. This supports that the t- and g-AFTERs are more robust than the L_1- and L_2-AFTERs.

3.6 Conclusions

Forecast combination is an important tool to achieve better forecasting accuracy when multiple candidate forecasters are available. Although many popular forecast combination methods do not necessarily exclude heavy tailed situations, little is found in the literature that examines the performances of forecast combination methods in such situations with theoretical characterizations.

In this chapter, we propose combination methods designed for cases when forecast errors exhibit heavy tail behaviors that can be modeled by a scaled Student's t-distribution and for the cases when the heaviness of the forecast errors is not easy to identify. The t-AFTER models the heavy-tailed random errors with scaled Student's t-distributions with unknown (or known) degrees of freedom and scale parameters. A candidate pool of degrees of freedom are proposed to solve the estimation problem and the resulting t-AFTER works well as seen in simulation and real example analysis.

However, in many cases the heaviness of the tails of the random errors is difficult to identify. Therefore, we design a combination process for general use and call it g-AFTER. For these situations, instead of assuming a certain distribution form for the random errors, a set of possible heaviness of the tails are considered and the combination process automatically decides which ones are more reasonable by giving them high weights. The numerical results suggest the performance of the g-AFTER is more robust than other popular combination methods because of its adaptive capability. The design of the g-AFTER provides a general idea: when there are multiple reasonable candidate distributions for the random errors, combining them in an AFTER scheme like the g-AFTER for forecast combination should work well.

3.7 Proofs of Theorems 3 and 4

3.7.1 Some Useful Simple Facts

These facts are used in the next subsection.

- Fact 1: $1-(1-t)^a \le \dfrac{at}{1-t}$ for $a \ge 0, 0 \le t < 1$. Let $f(t,a) = 1-(1-t)^a - at/(1-t)$, then $f(t,a) \le 0$ since $\partial f/\partial t = a(1-t)^{-2}((1-t)^{a+1}-1) \le 0$ and $f(0,a) = 0$.

- Fact 2: $\log(x) \le x-1$ for $x \ge 0$.

- Fact 3: For any $c > 0$, $B(a,b)/B(a,b+c)$ decreases as b increases. The proof is pure arithmetics and the key point is using the fact that $B(x,y) = \frac{x+y}{xy}\prod_{n=1}^{\infty}\left(1+\dfrac{xy}{n(x+y+n)}\right)^{-1}$.

- Fact 4: $E(1+\frac{Y^2}{\nu})^{-1} = \nu/(\nu+1)$, where $Y \sim t_\nu$ conditional on ν. Let $Z = Y\sqrt{(\nu+2)/\nu}$, then it is easy to show that $E(1+\frac{Y^2}{\nu})^{-1} = B(1/2,(\nu+2)/2)/B(1/2,\nu/2) = \nu/(\nu+1)$.

- Fact 5: $(s^2-1)/2 - \log(s) \le \frac{s_0+2}{2s_0}(1-s)^2$ if $s \ge s_0 > 0$. Using fact 2 to show that $-\log(s) = \log(1+(1-s)/s) \le (1-s)/s$.

3.7.2 Lemmas for the Proof of Theorem 3

Lemma 1 Let $h_\nu(x)$ be the density function of t_ν, $\underline{\nu} > 0$ and $\lambda > 0$ be constants. Then for any $0 < s_0 \le s$, $\underline{\nu} \le \min(\nu,\nu') - 2 \le \bar{\nu}$ and $|\nu - \nu'| \le \lambda$, we have

$$\int h_\nu(x)\log\frac{h_\nu(x)}{\frac{1}{s}h_{\nu'}\left(\frac{x-t}{s}\right)} \le C_1(1-s)^2 + C_2t^2 + C_3\left|\frac{\nu'-\nu}{\nu}\right|,$$

where C_1, C_2 and C_3 are constants depending on s_0, $\underline{\nu}$, $\bar{\nu}$ and λ.

Proof: After a proper reorganization, we have

$$E\log\frac{h_\nu(X)}{\frac{1}{s}h_{\nu'}\big(\frac{X-t}{s}\big)} = \log(s) + \frac{1}{2}\log\frac{\nu'}{\nu} + \log\frac{B(\frac{1}{2},\frac{\nu'}{2})}{B(\frac{1}{2},\frac{\nu}{2})} + E\left(\frac{1+\nu'}{2}\log\big(1+\frac{(X-t)^2}{s^2\nu'}\big) - \frac{1+\nu}{2}\log\frac{X^2+\nu}{\nu}\right)$$

- Let $\nu^* = \min(\nu,\nu')$ and using the Facts 1, 2 and 3, then:

$$\begin{aligned}
&\log\frac{B(\frac{1}{2},\frac{\nu'}{2})}{B(\frac{1}{2},\frac{\nu}{2})} \le \frac{|B(\frac{1}{2},\frac{\nu}{2}) - B(\frac{1}{2},\frac{\nu'}{2})|}{B(\frac{1}{2},\frac{\nu}{2})} \\
&= \frac{\int t^{-1/2}(1-t)^{\nu^*/2-1}(1-(1-t)^{|\nu-\nu'|/2})dt}{B(\frac{1}{2},\frac{\nu}{2})} \le \frac{\frac{|\nu-\nu'|}{2}\int t^{1/2}(1-t)^{\nu^*/2-2}dt}{B(\frac{1}{2},\frac{\nu}{2})} \\
&= \frac{|\nu-\nu'|}{2}\frac{B(\frac{3}{2},\frac{\nu^*-2}{2})}{B(\frac{1}{2},\frac{\nu}{2})} = \frac{|\nu-\nu'|}{2}\frac{B(\frac{3}{2},\frac{\nu^*-2}{2})}{B(\frac{1}{2},\frac{\nu^*-2}{2})}\frac{B(\frac{1}{2},\frac{\nu^*-2}{2})}{B(\frac{1}{2},\frac{\nu}{2})} \\
&= \frac{|\nu-\nu'|}{2}\frac{1}{\nu^*-1}\frac{B(\frac{1}{2},\frac{\nu}{2})}{B(\frac{1}{2},\frac{\nu^*+2}{2})} = \frac{|\nu-\nu'|}{\nu}\frac{\nu}{\nu^*-1}\frac{B(\frac{1}{2},\frac{\nu}{2})}{B(\frac{1}{2},\frac{\nu^*+2}{2})} \\
&\le \frac{|\nu-\nu'|}{\nu}\frac{\underline{\nu}+\lambda}{\underline{\nu}+1}\frac{B(\frac{1}{2},\frac{\nu}{2})}{B(\frac{1}{2},\frac{\nu^*+2}{2})} \le \frac{|\nu-\nu'|}{\nu}\frac{\underline{\nu}+\lambda}{\underline{\nu}+1}
\end{aligned}$$

- Using Fact 2 in A.1, it follows: $\frac{1}{2}\log\frac{\nu'}{\nu} \le \frac{1}{2}\frac{\nu'-\nu}{\nu} \le \frac{1}{2}\frac{|\nu'-\nu|}{\nu}$.

- It is easy to show that:

$$\begin{aligned}
&E\left\{\log(s) + \frac{1+\nu'}{2}\log\big(1+\frac{(X-t)^2}{s^2\nu'}\big) - \frac{1+\nu}{2}\log\big(1+\frac{X^2}{\nu}\big)\right\} \\
&= E\left\{\log(s) - (1+\nu')\log(s) + \frac{1+\nu'}{2}\log\left(\frac{s^2+\frac{(X-t)^2}{\nu'}}{1+\frac{X^2}{\nu}}\right) + \frac{\nu'-\nu}{2}\log\big(1+X^2/\nu\big)\right\} \\
&\le -\nu'\log(s) + E\left\{\frac{1+\nu'}{2}\frac{s^2-1+(X-t)^2/\nu'-X^2/\nu}{1+X^2/\nu} + X^2|\nu'-\nu|/\nu\right\} \\
&\le (2+\bar{\nu})\frac{2+s_0}{2s_0}(1-s)^2 + \frac{\underline{\nu}+3}{\underline{\nu}+2}t^2 + C_3^*\frac{|\nu'-\nu|}{\nu},
\end{aligned}$$

where C_3^* is a constant depending on s_0, $\underline{\nu}$, $\bar{\nu}$ and λ.

The proof can be completed by combining these steps.

Note that if ν is known, then $\nu = \nu'$. Then,

$$E \log \frac{h_\nu(X)}{\frac{1}{s} h_{\nu'}\left(\frac{X-t}{s}\right)} \leq \nu \frac{2+s_0}{2s_0}(1-s)^2 + \frac{1}{2}t^2.$$

Lemma 2 Let $h(x)$ be the density function of a double-exponential distribution with $\mu = 0$ and $d = 1$, then for $s_0 > 0$ and $s \geq s_0$ it follows:

$$\int h(x) \log \frac{h(x)}{\frac{1}{s} h\left(\frac{x-t}{s}\right)} \leq C_4(1-s)^2 + C_5 t^2,$$

where C_4 and C_5 are constants depending only on s_0.

Proof: since $h(y) = \frac{1}{2}\exp(-|y|)$ and $\exp(-x) \leq 1 - x + \frac{x^2}{2}$ for $x \geq 0$, then

$$\begin{aligned} E \log \frac{h(Y)}{\frac{1}{s} h\left(\frac{Y-t}{s}\right)} dy &= \log(s) + E\left(\frac{|Y-t|}{s}\right) - E|Y| = \log(s) + \frac{\exp(-t)+t}{s} - 1 \\ &\leq (s-1) + \frac{1+t^2/2}{s} - 1 = \frac{t^2}{2s} + (1-s)^2 \frac{1}{s} \\ &\leq \frac{t^2}{2s_0} + \frac{1}{s_0}(1-s)^2. \end{aligned}$$

Lemma 3 Let $h(y)$ be the density function of a standard normal distribution, then for $s_0 > 0$ and $s \geq s_0$ it follows:

$$\int h(x) \log \frac{h(x)}{\frac{1}{s} h\left(\frac{x-t}{s}\right)} \leq C_6(1-s)^2 + C_7 t^2,$$

where C_6 and C_7 are constants depending only on s_0.

Proof: using Fact 2,

$$\begin{aligned} E\log\frac{h(Y)}{\frac{1}{s}h\big(\frac{Y-t}{s}\big)}dy &= \log(s)+\frac{1+t^2-s^2}{2s^2} = \frac{1}{2s^2}t^2+\log(s)+\frac{1-s^2}{2s^2} \\ &\le \frac{1}{2s^2}t^2+(s-1)+\frac{1-s^2}{2s^2} = \frac{1}{2s^2}t^2+\frac{2s+1}{2s^2}(s-1)^2 \\ &\le \frac{1}{2s_0^2}t^2+\frac{2s_0+1}{2s_0^2}(s-1)^2. \end{aligned}$$

3.7.3 Proof of Theorem 3

Conditional on the information available until time point i, it is assumed that $\frac{Y_i-m_i}{s_i}\sim t_\nu$, where s_i is the conditional scale parameter at time i. Let $\hat{s}_{i,j}$ be the estimator of s_i from the j-th forecaster.

Let $f^n=\prod_{i=i_0+1}^{i_0+n}\frac{1}{s_i}h\left(\frac{y_i-m_i}{s_i}\right)$ and $q^n=\sum_{j=1}^K\pi_j\prod_{i=i_0+1}^{i_0+n}\frac{1}{\hat{s}_{i,j}}h\left(\frac{y_i-\hat{y}_{i,j}}{\hat{s}_{i,j}}\right)$, where $h(\cdot)$ is the density function of t_ν and π_j is the initial combining weight of the j-th forecaster. So, q^n is the estimator of f^n.

Then, for any $1\le j'\le J$,

$$\log(f^n/q^n)\le\log\frac{\prod_{i=i_0+1}^{i_0+n}\frac{1}{s_i}h\big(\frac{y_i-m_i}{s_i}\big)}{\pi_{j'}\prod_{i=i_0+1}^{i_0+n}\frac{1}{\hat{s}_{i,j'}}h\big(\frac{y_i-\hat{y}_{i,j'}}{\hat{s}_{i,j'}}\big)}=\log\frac{1}{\pi_{j'}}+\sum_{i=i_0+1}^{i_0+n}\log\frac{\frac{1}{s_i}h\big(\frac{y_i-m_i}{s_i}\big)}{\frac{1}{\hat{s}_{i,j'}}h\big(\frac{y_i-\hat{y}_{i,j'}}{\hat{s}_{i,j'}}\big)}$$

Conditional on all the information before time point i,

$$\begin{aligned} E_i\log\frac{\frac{1}{s_i}h\big(\frac{Y_i-m_i}{s_i}\big)}{\frac{1}{\hat{s}_{i,j'}}h\big(\frac{Y_i-\hat{y}_{i,j'}}{\hat{s}_{i,j'}}\big)} &= \int\frac{1}{s_i}h\big(\frac{y_i-m_i}{s_i}\big)\log\frac{\frac{1}{s_i}h\big(\frac{y_i-m_i}{s_i}\big)}{\frac{1}{\hat{s}_{i,j'}}h\big(\frac{y_i-\hat{y}_{i,j'}}{\hat{s}_{i,j'}}\big)}dy_i \\ &= \int h(x)\log\frac{h(x)}{\frac{1}{\hat{s}_{i,j'}/s_i}h\big(\frac{x-(\hat{y}_{i,j'}-m_i)/s_i}{\hat{s}_{i,j'}/s_i}\big)}dx \end{aligned}$$

By the Lemma 1 in A.2,

$$E_i \log \frac{\frac{1}{s_i} h\left(\frac{Y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{i,j'}} h\left(\frac{Y_i - \hat{y}_{i,j'}}{\hat{s}_{i,j'}}\right)} \leq \frac{(\hat{y}_{i,j'} - m_i)^2}{2s_i^2} + B_1 \frac{(\hat{s}_{i,j'} - s_i)^2}{s_i^2}$$

where $B_1 = \nu \frac{2+s_0}{2s_0}$. So,

$$\frac{1}{n} \sum_{i=i_0+1}^{i_0+n} ED(q_i || \hat{q}_i^{A_t}) \leq \inf_{1 \leq j \leq J} \left(\frac{\log \frac{1}{w_j^{A_t}}}{n} + \frac{1}{n} \sum_{i=i_0+1}^{i_0+n} E \frac{(\hat{y}_{i,j} - m_i)^2}{2s_i^2} + \frac{B_1}{n} \sum_{i=i_0+1}^{i_0+n} E \frac{(\hat{s}_{i,j} - s_i)^2}{s_i^2} \right)$$

From the Theorem 1 of Yang (2004), there exists a constant C depending on the parameters in Conditions 4 and 5′, such that,

$$ED(q_i || \hat{q}_i^{A_t}) \geq \frac{1}{C} E \frac{(m_i - \hat{y}_i^{A_t})^2}{\sigma_i^2}.$$

Therefore,

$$\frac{1}{n} \sum_{i=i_0+1}^{i_0+n} E \frac{(m_i - \hat{y}_i^{A_t})^2}{\sigma_i^2} \leq C \inf_{1 \leq j \leq J} \left(\frac{\log \frac{1}{w_j^{A_t}}}{n} + \frac{B_2}{n} \sum_{i=i_0+1}^{i_0+n} E \frac{(\hat{y}_{i,j} - m_i)^2}{\sigma_i^2} + \frac{B_3}{n} \sum_{i=i_0+1}^{i_0+n} E \frac{(\hat{s}_{i,j} - s_i)^2}{s_i^2} \right),$$

where B_2 is a function of ν and B_3 is deducted the same as B_1 but under Condition 5′ instead of Condition 5.

3.7.4 Proof of Theorem 4

Essential part of the proof of Theorem 4 is provided in this subsection. We only provide the steps of the proof when the random errors are scaled Student's t-distributed since proof of other situations are similar.

Let $\hat{s}_{i,j,k}$ be the estimator of s_i from the j-th forecaster assuming ν_k is the true degrees of freedom. If Condition 4 holds, then obviously

$$q^n \geq \sum_{k=1}^{K}\sum_{j=1}^{J} c_2 w_{j,k}^{A_t}/G \prod_{i=i_0+1}^{i_0+n} \frac{1}{\hat{s}_{i,j,k}} h_{\nu_l}\left(\frac{y_i-\hat{y}_{i,j}}{\hat{s}_{i,j,k}}\right).$$

So, for any j^* and k^*,

$$\begin{aligned}\log\frac{f^n}{q^n} &\leq \log \frac{\prod_{i=i_0+1}^{i_0+n}\frac{1}{s_i}h\left(\frac{y_i-m_i}{s_i}\right)}{c_2 w_{j^*,k^*}^{A_t}/G\prod_{i=i_0+1}^{i_0+n}\frac{1}{\hat{s}_{i,j^*,k^*}}h_{\nu_{k^*}}\left(\frac{y_i-\hat{y}_{i,j^*}}{\hat{s}_{i,j^*,k^*}}\right)} \\ &= \log\frac{G}{c_2 w_{j^*,k^*}^{A_t}} + \sum_{i=i_0+1}^{i_0+n}\log\frac{\frac{1}{s_i}h\left(\frac{y_i-m_i}{s_i}\right)}{\frac{1}{\hat{s}_{i,j^*,k^*}}h\left(\frac{y_i-\hat{y}_{i,j^*}}{\hat{s}_{i,j^*,k^*}}\right)}.\end{aligned}$$

Similarly, by the Lemma 1 in A.2,

$$E_i\log\frac{\frac{1}{s_i}h\left(\frac{Y_i-m_i}{s_i}\right)}{\frac{1}{\hat{s}_{i,j^*,k^*}}h\left(\frac{Y_i-\hat{y}_{i,j^*}}{\hat{s}_{i,j^*,k^*}}\right)} \leq B_1\frac{(\hat{y}_{i,j^*}-m_i)^2}{\sigma_i^2} + B_2\frac{(\hat{s}_{i,j^*,k^*}-s_i)^2}{s_i^2} + B_3\left|\frac{\nu_k-\nu}{\nu}\right|.$$

The rest of the proof is similar to that of Theorem 3.

Chapter 4

R Package: AFTER

To ease the implementation and application of the new combination methods proposed in chapters 2 and 3 for potential users, we compiled an R package called `AFTER`. In this chapter, we will provide descriptions of the main functions in this package along with some examples.

4.1 Basic Description

- Type: Package
- Title: AFTER version 1.0
- Date: 2014-10-20
- Author: Gang Cheng, Yuhong Yang
- Maintainer: Gang Cheng, Yuhong Yang
- Description: The newly proposed AFTER methods discussed in this dissertation and many other popular forecast combination methods, including the regression-based methods and variance-covariance matrix estimation-based methods are implemented for time series data or other data frames.

- Depends: R (>2.10.1)
- License: GPL (≥2)
- NeedsCompilation: Yes
- Repository: CRAN

4.2 Main Functions

In the section, four main functions, `AFTER`, `LinRegComb`, and `BGComb`, are introduced with examples.

4.2.1 Function `AFTER`

AFTER *Combination methods in the AFTER family*

Description

The L_2-AFTER, L_1-AFTER, L_{210}-AFTER, t-AFTER and g-AFTER are implemented.

Usage

```
AFTER(trainCand, trainActu, Cand, method="L2", window = "ALL")
```

Arguments

`trainCand` It is the matrix of candidate forecasters. Each row is a vector of forecasts for the same target variable.

`trainActu` It is a vector of values. The i-th element is the observed value of the variable that the i-th row of `trainCand` forecasts.

`Cand` It can be either a matrix or a vector of candidate forecasts. If it is a matrix, then each row is a vector of forecasts for the same value.

`method` The default method is the L_2-AFTER. If the user wants to implement the L_1-AFTER, put `method` = "L1"; if L_{210}-AFTER, put `method` ="L210" if you use the default setting ($\alpha_1 = -\alpha_2 = 2$); put `method` ="L210_(alpha1,alpha2)" to specify the options for α_1 and α_2. To implement the t-AFTER, put `method` ="t" for default t-AFTER or for example, `method` = "t_(1,3,5,7)" to call t-AFTER with $(1, 3, 5, 7)$ as the candidate degrees of freedom pool. For g-AFTER, put `method` = "g" for the default setting and `g_(1,4,6,7)` to call g-AFTER with the t-AFTER with $(1, 3, 5, 7)$ as the candidate degrees of freedom pool.

`window` It can be either "ALL "or a positive integer. If "ALL ", then all data in `trainCand` and `trainActu` are used to calculate the combination weights; if it is a positive integer (say k) , then the last k rows of `trainCand` and k elements of `trainActu` are used.

Value(s)

This function returns the combined forecast(s) from the combination methods specified.

Details

When the random errors in the underlying true model are roughly normally distributed, the L_2-AFTER is proper; the L_1-AFTER is more appropriate when there are occasional outliers. If the goal to control the number of large forecast errors, the L_{210}-AFTER is a reasonable choice. When random errors are considered (suspected) to follow heavy tailed distributions, then the t-AFTER (g-AFTER) is proper.

In the t-AFTER (and g-AFTER), to estimate the scale parameter of a scaled student's t-distribution for each individual candidate forecaster, we follow the steps below:

1. Get the historical forecast errors.
2. Get the median of the absolute forecast errors.
3. The estimate of the scale parameter is the median from step 2 divided by the theoretical median of the standard student's t-distribution with the given degrees of freedom.

Examples

```
### generate input data
> trainCand = matrix(rnorm(100),25,4);
> trianActu = rowMeans(trainCand) + rnorm(25,0,0.2);
> Cand = rnorm(4);

### default method is the L2-AFTER using all historical data
>AFTER(trainCand, trainActu, Cand)

### t-AFTER with (1,10,20) as candidate degrees of freedom
```

```
### using the most recent 20 observations
> Cand = matrix(rnorm(12),3,4);
> AFTER(trainCand, trainActu, Cand, method="t_(1,10,20)", window = 20)

### L210-AFTER with (alpha1=1,alpha2=-3)
### use the most recent 20 observations
 > AFTER(trainCand, trainActu, Cand, method="L210_(1,-3)", window = 20)
```

4.2.2 Function `LinRegComb`

LinRegComb *Combination methods based on linear regressions*

Description

This function can generate combined forecasts from linear regression with and without constraint and by the shrinkage method. See section 1.2 for details. Here, we only handle the cases that the number of forecast models is smaller than the number of forecast periods/points.

Usage

```
LinRegComb(trainCand, trainActu, Cand, contraint = "Y", shrink = "N",
           window = "ALL")
```

Arguments

The `trainCand`, `trainActu`, `Cand` and `window` have the same definition as those for function `AFTER`.

`constraint` It can only be `"Y"` or `"N"`. If `"Y"`, then the weights are from constrained linear regression; otherwise, ordinary linear regression is called.

`shrink` It can only be N or a value in between 0 and 1. It is only called when constraint = `"Y"`. When shrink = `"N"`, then the shrinkage method is not called; otherwise the ordinary constrained linear regression is mixed with simple average with mix weight extracted in `shrink`.

Value(s)

This function returns the combined forecast(s) from the combination methods specified.

Details

Constrained linear regression based combination is usually used as an alternative to the linear regression based combination when there are occasional outliers in the forecasts because the associated combination weights are relatively more stable. The shrinkage method is more stable then a constrained linear regression-based combination when outliers are present, but it may sacrifice some prediction accuracy.

Examples

```
### generate input data
> trainCand = matrix(rnorm(100),25,4);
> trianActu = rowMeans(trainCand) + rnorm(25,0,0.2);
> Cand = rnorm(4);
```

```
### default method is the ordinary linear regression method
> LinRegComb(trainCand, trainActu, Cand);

### put constraint and shrinkage
> Cand = matrix(rnorm(12),3,4);
> LinRegComb(trainCand, trainActu, Cand, contraint = N, shrink = 0.9, window=15);
```

4.2.3 Function `BGComb`

BGComb *Combination methods based on Bates and Granger (1969)*

Description

Three versions of forecast combination methods based on the estimation of covariance matrix of candidate forecasts are implemented here. See section 1.2 for details.

Usage

```
BGComb(trainCand, trainActu, Cand, window="ALL", discount = 1)
```

Arguments

The `trainCand`, `trainActu`, `Cand` and `window` have the same definition as those for function `AFTER`.

`discount` It can only be a value in between 0 and 1.

Value(s)

This function returns the combined forecast(s) from the combination methods specified.

Details

If the user wants to put more credibility to more recent observations in a time series setting, then using a short rolling window or a discount factor that is small serves the goal well. If the historical data are not too much different, then using all data is more reasonable.

Examples

```
### generate input data
> trainCand = matrix(rnorm(100),25,4);
> trianActu = rowMeans(trainCand) + rnorm(25,0,0.2);
> Cand = rnorm(4);

### default method is BG using all historical data without discount factor
>BGComb(trainCand, trainActu, Cand);

### specify different data window with a predetermined discount factor
> Cand = matrix(rnorm(12),3,4);
> BGComb(trainCand, trainActu, Cand, window= 20, discount = 0.9);
```

Chapter 5

Future Work

5.1 Combination for Improvement via AFTER

L_2-AFTER from Yang (2004), L_1-AFTER from Wei & Yang (2012), L_{210}-AFTER from Cheng & Yang (2014) and the t- and g-AFTER from this dissertation are proposed for adapting the performance of the best individual candidate forecasts in any given time periods. However, as discussed in section 1.3, if all candidate forecasts are not acceptable in terms of forecast accuracy, then performing as the best of them can still be not acceptable for real applications. To the best of our knowledge, no forecast combination method was found in literature that was proposed to handle the "weak candidate forecast" situations.

We propose a new combination frame that can provide competitive combined forecasts when the best ones of the individual candidate forecasts are not quite acceptable in terms of prediction accuracies. This frame automatically and dynamically detects which goal, adaptation versus improvement, is more proper given the set of historical data available and then directs the combination procedure to the right flow.

Specifically, there are three major steps of this frame:

1. Try to generate new and potentially better forecasts using the given candidate forecasts and the related actual values. This step can be broken into three

sub-steps:

(a) Split the historical data into two parts: The first part is called training data set and the second is called testing data set. On the training data set, train statistical methods, such as regression and boosting, that have potentials to generate "stronger" forecasts. Note that multiple methods can be considered in this phase.

(b) Use the trained methods to make predictions on the testing data set. The methods used to make the predictions can be from the training data set or dynamically updated through the testing data set.

(c) Combine all the candidate forecasts available, including the original candidate forecasts and the newly generate ones, via AFTER methods. Note that, if the original forecasts are consistently weaker than the newly generated candidates, then the original forecasts are not put into AFTER for generating the final forecasts.

If the original candidate forecasts are weak, then there is room for improvement if the statistical procedures picked to generate the strong ones are properly specified. So this new AFTER frame can adapt the performance of the best "stronger" forecasts instead of the "weak" ones. Some preliminary numerical and theoretical works are done and the new method supports our goal well. This paper will be completed later this year.

5.2 A Note to the "Forecast Combination Puzzle"

Simple average (`SA`) and other similar simple methods are widely found to be competitive compared to other relatively more sophisticated methods, such as `BG`, in literature. This phenomenon is usually referred to as the"forecast combination puzzle".

See, e.g., Smith & Wallis (2009) for details.

However, there is an obvious conflict in literature: On one side, many papers which propose new combination methods often used `SA` as a benchmark and demonstrate the advantages of the their methods by showing that they beat `SA` numerically. See, Wei & Yang (2012) and Hsiao & Wan (2014), for examples. However, on the other side, as Smith & Wallis (2009) discussed, simple methods such as `SA` are found to outperform sophisticated methods. For example, Stock & Watson (2003) and Stock & Watson (2004), showed that the `SA` outperforms regression based methods significantly based on the empirical study on some real data sets.

In our paper, we will address why different works have such conflicting conclusions. There are three major parts of the paper: 1). A comprehensive review of the literature showing the how the related works obtained their conclusions about the "forecast combination puzzle"; 2). Showing that at least, a modified AFTER frame can beat `SA` consistently in almost all the numerical examples discussed in literature. So, there is no "forecast combination puzzle"; 3). Explain why works in literature have different conclusions.

Chapter 6

Conclusion and Discussion

In general, to serve specific forecast goals, for example to have a small mean square forecast error or to have fewer large forecast errors in a given evaluation period, one can pick a relatively more proper combination method from the pool of existing methods. Another way is to design a related new combination method. In this dissertation, we designed new forecast combination methods for two important practical forecast problems.

Besides the popular predictive accuracy measures, such as the mean square forecast errors and the mean absolute forecast errors, the frequency of the forecast errors that is larger than the pre-determined tolerance in magnitude can be also very important to forecast service users in many areas. The control of the frequency of large forecast errors is more critical when the candidate forecasts have occasional large forecast errors. If occasional large forecast errors are present, a robust forecast combination method such as L_1-AFTER can be applied. However, we have noticed a somehow unnoticed phenomenon that although robust forecast combination methods such as the L_1-AFTER can provide overall more accurate forecasts, its forecasts may have more large forecast errors. So there is a need for a balance between robustness and large forecast error protection. In chapter 2, we first propose a new loss function, the L_{210}-loss which serves three goals simultaneously: 1). It penalizes the number of

large forecast errors; 2). It is a balance between robustness and outlier protection; 3). Using it in AFTER can achieve some nice theoretical risk bound for the related methods. The L_{210}-AFTER serves our goals nicely as seen in systematic numerical examples.

When we go further, what if the frequency of large forecast errors in candidate forecasts is significant instead of occasional? Can we find or design a proper combination method for that? Actually, to the best of our knowledge, no work in literature discusses effective forecast combination when the random errors in the true models have heavy tails, which usually leads to a significant amount of large forecast errors in candidate forecasts. So instead of using some flat-tailed distributions to model the random errors, we use scaled students' t-distributions with low degrees of freedom. When incorporating the scaled students' t-distributions assumption into the AFTER frame, there is an issue of estimating the scale parameter and degrees of freedom simultaneously. So we proposed a two-step procedure to estimate them which firstly decides a candidate pool of degrees of freedom and then estimates the scale parameter for each candidate degree of freedom for each candidate individual forecast. Then the AFTER procedure can estimate the likelihood of each degrees of freedom and scale parameter combination. The proposed t-AFTER and g-AFTER work well as demonstrated in numerical examples.

The two projects in chapters 2 and 3 inspired two followup projects. The methods proposed in this dissertation are based on the AFTER frame for adaptation. How to make some simple modification to AFTER to combine for improvement is an interesting topic. Either the AFTERs for adaptation or the AFTERs for improvement are based on AFTER from Yang (2004), which is usually considered a sophisticated method. Based on the "forecast combination puzzle", AFTER can often be outperformed by simple methods such as `SA`. However, this is not true from the numerical results in chapters 2 and 3. So, to figure out how the literature addresses "fore-

cast combination puzzle" and provide our insights should be an interesting topic to research. There two papers described in chapter 5 are under writing. If things go smoothly, the papers of these two projects will be submitted to journals soon.

References

Altavilla, C., De Grauwe, P. (2010) "Forecasting and combining competing models of exchange rate determination," *Applied Economics*, 42, 3455–3480.

Armstrong, J.S. (2007), "Significance Tests Harm Progress in Forecasting," *International Journal of Forecasting*, 23, 321–327.

Bates, J.M., Granger, C.W.J. (1969), "The Combination of Forecasts," *OR*, 20, 451–468.

Catoni, O. (1999), *'Universal' Aggregation Rules with Exact Bias Bound*, Preprint.

Catoni, O. (2004), *Statistical Learning Theory and Stochastic Optimization*, New York: Springer.

Chen, Z., Yang, Y. (2004), "Assessing Forecast Accuracy Measures," Preprint # 10, 2004, Department of Statistics, Iowa State University.

Chen, Z., Yang, Y. (2007), "Time Series Models for Forecasting: Testing or Combining," *Studies in Nonlinear Dynamics and Econometrics*, 11 (1), Article 3.

Cheng, G., Yang, Y. (2014), "Forecast Combination with Outlier Protection," *International Journal of Forecasting*, 10.1016/j.ijforecast.2014.06.004, Forthcoming.

Christoffersen, P., Diebold, F.X. (1997), "Optimal Prediction Under Asymmetrical Loss," *Econometric Theory*, 13, 806–817.

Clemen, R.T. (1989), "Combining Forecasts: a Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

Diebold, F.X. (2001), *Elements of Forecasting* (2nd ed.), South-Western Publishing.

Elliott, G., Timmermann, A. (2004), "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, 122, 47–49.

Fan, S., Chen, L. and Lee, W.J. (2008) "Short-term load forecasting using comprehensive combination based on multi-meteorological information," *Industrial and Commercial Power Systems Technical Conference*, ICPS, IEEE/IAS.

Fernandez, C. and Steel, M. F. J., (1999) "Multivariate Student-t regression models: Pitfalls and inference," *Biometrika*, 86 (1), 153–167.

Fonseca, T.C.O., Ferreira, M.A.R. and Migon, H. S. (2008) "Objective bayesian analysis for the Student-t regression model," *Biometrika*, 95, 325–333.

Glasserman, P., Heidelberger, P. and Shahabuddin, P. (2002) "Portfolio value-at-risk with heavy-tailed risk factors," *Mathematical Finance*, 12, 239–269.

Granger, C.W.J., Newbold, P. (1986), *Forecasting Economic Time Series* (2nd ed.), New York: Academic Press.

Granger, C.W.J., Pesaran, M.H. (2000), "Economic and Statistical Measures of Forecast Accuracy," *Journal of Forecasting*, 19, 537–560.

Granger, C.W.J., Ramanathan, R. (1984), "Improved Methods of Forecasting," *Journal of Forecasting*, 3, 197–204.

Hsiao, C., & Wan, S. K. (2014). "Is there an optimal forecast combination?," *Journal of Econometrics*, 178, 294-309.

Hansen, B.E. (2008), "Least Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350.

Harvey, A.C. (2013) "Dynamic models for volatility and heavy tails: With applications to financial and economical time series (pp. 69)," NYC, USA: Cambridge University Press.

Ing, C.K. (2007) "Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series," *Annals of Statistics* 35, 1238–1277.

Ing, C.K., Sin, C.-Y., and Yu, S.-H. (2012) "Model selection for integrated autoregressive processes of infinite order," *Journal of Multivariate Analysis*, 106, 57–71.

Inoue, A. and Kilian, L. (2008) "How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation," *Journal of the American Statistical Association*, 103 (482), 511–522.

Kan, R. and Zhou, G. (2003) Modeling non-normality using multivariate t: Implications for asset pricing. *Technical report*, Rotman School of Management, University of Toronto, Toronto, Canada.

Lahiri, K., Peng, H., Zhao, Y. (2013), "Machine learning and forecast combination in incomplete panels," *University at Albany, SUNY, Department of Economics in its series Discussion Papers*, 13–01.

Liu, Y. and Wu, Y. (2007), "Variable selection via a combination of the L0 and L1 penalties," *Journal of Computational and Graphical Statistics*, 16, 4, 782 – 798.

Makridakis, S., Hibon, M. (2000), "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16, 451–476.

Marinelli, C., Rachev, S. and Roll, R. (2001) "Subordinated exchange rate models: Evidence for heavy tailed distributions and long-range dependence," *Mathematical and Computer Modelling*, 34, 955–1001.

Min, C.K., Zellner, A. (1993), "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.

Nesterov, Y. (2004), "Introductory Lectures on Convex Optimization: A Basic Course," *Kluwer Academic Publishers*, 63–64.

Newbold, P., Harvey, D. I. (2002), "Forecast Combination and Encompassing," in A companion to economic forecasting, eds, Clemenets, M. P. and Hendry, D. F., Oxford: Blackwells.

Pai, P.F., Lin, C.S. (2005), "A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting," *Omega*, 33 (6), 497–505.

Sancetta, A. (2010) "Recursive forecast combination for dependent heterogeneous data," *Econometric theory*, 26, 598–631.

Sanchez, I. (2008) "Adaptive combination of forecasts with application to wind energy," *International Journal of Forecasting*, 24, 679–693.

Smith, J., & Wallis, K. F. (2009). "A simple explanation of the forecast combination puzzle," *Oxford Bulletin of Economics and Statistics*, 71(3), 331–355.

Stock, J.H., Watson, M.W. (1999), "Forecasting inflation," *Journal of Monetary Economics*, 44, 293–335.

Stock, J.H., Watson, M.W. (2003), "Forecasting Output and Inflation: the Role of Asset Prices," *Journal of Economic Literature*, 41, 788–829.

Stock, J. H., & Watson, M. W. (2004). "Combination forecasts of output growth in a seven country data set," *Journal of Forecasting*, 23(6), 405-430.

Stock, J.H. and Watson, M.W. (2006) "Forecasting with many predictors," *Handbook of economic forecasting*, 1, 515-554.

Timmermann, A. (2000), "Density Forecasting in Economics and Finance," *Journal of Forecasting*, 19, 231–234.

Timmermann, A. (2006), "Forecast Combinations," in Handbook of Economic Forecasting, eds, Elliott, G., Granger, C.W.J., Timmermann, A., Amsterdam: Elsevier.

T sybakov, A. B. (2003), "Optimal rates of aggregation. Learning Theory and Kernel Machines," *Lecture Notes in Artificial Intelligence*, 2777, 303–313. Heidelberg: Springer.

Wei, X., Yang, Y. (2012), "Robust Forecast Combinations," *Journal of Econometrics*, 166, 224–236.

West, K.D., Edison, H.J., Cho, D. (1997), "A Utility Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics*, 35, 23–46.

Yang, Y. (2000), "Mixing Strategies for Density Estimation," *Annals of Statistics*, 28, 75–87.

Yang, Y. (2004), "Combining Forecasting Procedures: Some Theoretical Results," *Econometric Theory*, 20, 176–222.

Zellner, A. (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of the American Statistical Association*, 81, 446–451.

Zeng, T., Swanson, N.R. (1998), "Predictive Evaluation of Econometric Forecasting Models in Commodity Futures Markets," *Studies in Nonlinear Dynamics and Econometrics, Berkeley Electronic Press*, 2 (4), 6.

Zhang, C.H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The annals of statistics*, 38, 894–942.

Zhang, X., Lu, Z. and Zou, G. (2013) "Adaptively combined forecasting for discrete response time series," *Journal of Econometrics*, 176 (1), 80–91.

Zou, H., Yang, Y. (2004), "Combining Time Series Models for Forecasting," *International journal of Forecasting*, 20, 69–84.