Cheng, Gang; Wang, Sicong; Yang, Yuhong

Article

Forecast combination under heavy-tailed errors

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Cheng, Gang; Wang, Sicong; Yang, Yuhong (2015) : Forecast combination under heavy-tailed errors, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 3, Iss. 4, pp. 797-824, https://doi.org/10.3390/econometrics3040797

Econometrics **2015**, *3*, 797-824; doi:10.3390/econometrics3040797

OPEN ACCESS

econometrics

ISSN 2225-1146

www.mdpi.com/journal/econometrics

Article

Forecast Combination under Heavy-Tailed Errors

Gang Cheng *, Sicong Wang and Yuhong Yang

School of Statistics, University of Minnesota at Twin Cities, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455, USA; E-Mails: huihui3119@gmail.com (S.W.); yangx374@umn.edu (Y.Y.)

* Author to whom correspondence should be addressed; E-Mail: chen2285@umn.edu; Tel.: +1-612-508-5360.

Academic Editor: Isabel Casas

Received: 22 August 2015 / Accepted: 10 November 2015 / Published: 23 November 2015

Abstract: Forecast combination has been proven to be a very important technique to obtain accurate predictions for various applications in economics, finance, marketing and many other areas. In many applications, forecast errors exhibit heavy-tailed behaviors for various reasons. Unfortunately, to our knowledge, little has been done to obtain reliable forecast combinations for such situations. The familiar forecast combination methods, such as simple average, least squares regression or those based on the variance-covariance of the forecasts, may perform very poorly due to the fact that outliers tend to occur, and they make these methods have unstable weights, leading to un-robust forecasts. To address this problem, in this paper, we propose two nonparametric forecast combination methods. One is specially proposed for the situations in which 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 a shortage of data and/or an evolving data-generating process. Adaptive risk bounds of both methods are developed. They show that the resulting combined forecasts yield near optimal mean forecast errors relative to the candidate forecasts. Simulations and a real example demonstrate their superior performance in that they indeed tend to have significantly smaller prediction errors than the previous combination methods in the presence of forecast outliers.

Keywords: forecast combination; heavy tails; robustness; time series models; nonparametric forecast combination

JEL classifications: C40; C51; C53

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 the applications of forecast combinations in areas, such as economics, finance, tourism and wind power generation in the last fifty years.

Many combination methods have been proposed from different perspectives since the seminal work of forecast combination by Bates & Granger [1]. See the discussions and summaries in Clemen [2], Newbold & Harvey [3] and Timmermann [4] for key developments and many references. More recently, Lahiri *et al.* [5] provide theoretical and numerical comparisons between adaptive and simple forecast combination methods; Armstrong, Green and Graefe [6] propose important principles to follow, centered on the golden rule of being conservative, for building accurate forecasts, and verify them empirically based on an examination of previously-published studies; Green & Armstrong [7] review studies that compare simple and complicated methods and conclude that complexity actually substantially increases forecast error. They advocate the use of sophisticatedly simple methods instead of complicated ones that are hard to understand. This is in line with the fact that complicated methods often incur unnecessarily larger instability and variability in prediction (see, e.g., Subsection 3.1 of Yang [8]). While it seems clear that researchers agree that forecast combination is very useful, they differ in their opinions on how to do forecast combination properly. Needless to say, there are many possibly drastically different scenarios one can envision for the problem of forecast combination in terms of the accuracy of the candidate forecasts, their relationships, the structure changes, the characteristics of the forecast errors and more, which naturally favor different methods to be top performers. Therefore, the availability of many combination methods and disputes on their rankings and merits, in our view, are not only expected, but also helpful to collectively reach a better understanding of the key issues in the research area by further rigorous theoretical and empirical investigations.

The present work concerns forecast combination when the forecast errors exhibit heavy-tailed behaviors, which means that the decay of the probability density function (or an estimate) of the forecast errors is much slower than that of the normal distribution. To our knowledge, few studies have proposed/discussed forecast combination methods that target such situations, where the familiar forecast combination methods, such as simple average, least squares regression with or without constraints or those based on the variance-covariance of the forecasts, may perform very poorly (some numerical examples are provided in Sections 4 and 5 in this paper).

Heavy-tailed behaviors of forecast errors may come from different sources. First, many important variables in finance, economics and other areas are known to have heavy tails. For example, currency exchange rates have long been believed to have heavy-tailed behaviors, and Marinelli *et al.* [9], for instance, discussed the evidences of heavy-tailed distributions to model them. Some key macroeconomic indices, such as GDP, are also believed to have heavy-tailed tendencies, and Harvey [10], for instance, modeled the U.S. GDP with Student's t-distributions with low degrees of freedom. The heavy tails of

the variables to be forecast naturally tend to cause heavy-tailed behaviors of the forecast errors. Second, even if the target variables themselves have light tails, the variables in the information set may have long tails for various reasons, which can induce heavy tails of the forecast errors. Third, for a difficult target variable, we may also observe heavy-tailed forecast errors from predictive models when data available for model training are limited, even if the true data-generating processes have relatively normal tails.

Clearly, when some of the forecast errors of the candidate forecasters are unusually large, if a forecast combination method does not take it into consideration, the final forecast may even fully inherit large prediction errors, which may then have severe practical consequences on decisions based on the forecast. Therefore, it is crucial to devise combination methods that can deal with heavy-tailed forecast errors for robust and reliable final performances. In the rest of the work, for convenience, heavy-tailed distributions may sometimes loosely refer to distributions with tails heavier than Gaussian distributions, although specific choices, such as scaled t-distributions, will be studied.

In this paper, we propose two forecast combination methods. 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 (see below). The other is designed for more general uses. The design of these two methods follows the spirit of the adaptive forecasting through exponential re-weighting (AFTER) combination scheme by Yang [8]. The idea of the AFTER scheme is that the exponentiated cumulative historical performances of the candidate forecasts are informative and can be used to assign their combination weights for the future. This way of using the historical performances of the candidate forecasts for weighting has a natural tie to information theory and provides a near optimal final performance in mean forecasting errors. For example, if the random errors in the true model are from a normal distribution, then the weight of a candidate forecast by AFTER is proportional to $\exp(-L_2)$, where L_2 is the cumulative historical mean squared forecast error of the forecast. For the fist method mentioned above, we assume that the forecast errors follow a scaled Student's t-distribution with a possibly unknown scale parameter and degrees of freedom. Note that if a random variable X satisfies that $X/s \sim t_\nu$ for some $s > 0$, where t_ν is a standard t-distribution with degrees of freedom ν, we say X has a t_ν distribution with scale parameter s. 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 for the second method. In either case, no parametric assumptions are needed on the relationships of the candidate forecasts.

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 of the candidate forecasters, because the estimation of the conditional scale parameters is straightforward (see, e.g., Zou & Yang [11] and Wei & Yang [12], for more details). However, this is not true if a scaled 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 [13] and Fonseca *et al.* [14] 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 model suffer from monotonic likelihood because the likelihood goes to

infinity as the scale parameter goes to zero if the degrees of freedom ν are not large enough. To deal with this difficulty, methods other than the 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 the method of moments or other tools (see, e.g., Kan & Zhou [15]).

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 mixture of all of the candidate density estimates using the weights. More details about this procedure, including how to determine the pool of candidate estimates, are available in Section 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 degree of freedom 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 the variance-covariance of forecasters require the existence of variances. Regression-based forecast combination methods (see, e.g., Granger and Ramanathan [16]) 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 [17] assumed that the tails of the target variables are no heavier than exponential decays, which restricts the heaviness of the tails of the forecast errors. Wei & Yang [12] designed a method for errors heavier than the normal distributions, but not heavier than the double-exponential distributions. More recently, Cheng & Yang [18] advocate the incorporation of a smooth surrogate of the L_0-loss in the performance measure for weighting to reduce the occurrence of outlier forecasts. 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 performance of the proposed methods will be examined via simulations and a real data example. We consider two simulation settings, depending on the data-generating processes being from regression models or time series models. Several error distributions are used, and they have different degrees of heavy tails. The new methods are compared to earlier versions of AFTER, as well as some popular combination methods. Their performances in heavy-tailed situations are indeed better than the competitors and are still among or close to the best, even if the forecast errors have normal tails. For a real data application, we use 1428 time series variables from M3-competition data (see Makridakis & Hibon [19]). The M3-competition data are very popular in empirical studies in econometrics, machine learning and statistics to validate the performances of forecasting methods. For each of the variables in this dataset, forecast sequences based on 24 popular forecast methods are provided. The overall evaluation on the 1428 variables shows that our proposed methods, especially the one for general

purposes, compare favorably to others. To gain more insight, we pick out a subset of the 1428 variables that have heavy-tailed forecast errors, and it is seen that the the new methods behave nicely, as intended.

The plan of the paper is as follows: Section 2 introduces the forecast combination method designed for heavy-tailed error distributions. In Section 3, a more general combination method is proposed. Simulations are presented in Section 4, and Section 5 provides a real data example. Section 6 includes a brief concluding discussion. The proofs of the theoretical results are in the Appendix.

2. The t-AFTER Methodology

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

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 random 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 zero and scale parameter one.

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:

$$\hat{y}_i = \langle \hat{Y}_i, W_i \rangle, \tag{1}$$

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.

2.2. The Existing AFTER Methods: The L_2- and L_1-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 [8]. It has been applied and studied in, e.g., Fonseca *et al.* [14], Inoue & Kilian [20], Sanchez [21], Altavilla & De Grauwe [22] and Lahiri *et al.* [5] and Zhang *et al.* [23] 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:

$$W_i = \frac{\mathbf{l}_{i-1}}{||\mathbf{l}_{i-1}||_1}, \tag{2}$$

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

$$l_{i-1,j} = w_j \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j}} h\left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j}}\right), \tag{3}$$

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

Below, the most commonly-used AFTER procedures, the L_2-AFTER from Zou & Yang [11] and the L_1-AFTER from Wei & Yang [12], are briefly introduced.

L_2-AFTER: When the random 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 performance 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 (3), then plug the new $\mathbf{l}_{i-1}$ into (2). The $\hat{s}_{i,j}$ used in the L_2-AFTER, denoted as $\hat{\sigma}_{i,j}$, is the sample standard deviation of $\{y_{i'} - \hat{y}_{i',j}\}_{i'=1}^{i-1}$, assuming the random errors are independent and identically distributed.

L_1-AFTER: Let f_{DE} be the pdf of a double-exponential distribution with scale parameter one and location parameter zero. To get $W_i^{A_1}$, one can follow the same procedure for $W_i^{A_2}$, but use f_{DE} as the h in (3). The $\hat{s}_{i,j}$ used in the L_1-AFTER, denoted as $\hat{d}_{i,j}$, is the mean of $\{|y_{i'} - \hat{y}_{i',j}|\}_{i'=1}^{i-1}$. The L_1-AFTER method was designed for robust combination when the random errors have occasional outliers. See Wei and Yang [12] for details.

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:

1. We decide a pool of candidate degrees of freedom with size K. The elements in the pool are considered to be close to the degrees of freedom of the Students' t-distribution that describes the random errors well. For each element in the set, we assume it is the true degrees of freedom to estimate the related scale parameter. Therefore, we have K sets of estimate for the degrees of freedom and scale parameter pair.
2. For each of the K sets of the estimate, we find its probability to be the true one based on the relative historical performances.

This two-step procedure is used in the t-AFTER method for forecast combination when the random errors have heavy tails that can be described well by a Students' t-distribution.

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 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 steps:

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 $\hat{y}_i^{A_t}$:

$$W_i^{A_t} = \frac{\mathbf{l}_{i-1}^{A_t}}{||\mathbf{l}_{i-1}^{A_t}||_1}, \quad \hat{y}_i^{A_t} = \langle \hat{Y}_i, W_i^{A_t} \rangle, \tag{4}$$

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 $i \geq i_0 + 1$,

$$l_{i-1,j}^{A_t} = \sum_{k=1}^{K} l_{i-1,j,k}^{A_t} \quad \text{with} \quad l_{i-1,j,k}^{A_t} = w_{j,k} \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j,k}} f_t \left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j,k}} \middle| \nu_k \right), \tag{5}$$

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 datasets. 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 three or five. In this situation, from our experiences, $\Omega = \{1, 3\}$ or $\{1, 3, 5\}$ works well.

Obviously, when the random errors in the true model follow a scaled Student's t-distribution with a known degree of freedom ν, then $\Omega := \{\nu\}$. Then, (5) can be simplified into:

$$l_{i-1,j}^{A_t} = w_j \prod_{i' \geq i_0}^{i-1} \frac{1}{\hat{s}_{i',j}} f_t \left(\frac{y_{i'} - \hat{y}_{i',j}}{\hat{s}_{i',j}} \middle| \nu \right), \tag{6}$$

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.

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 (the third remarks of Theorem 2).

2.4.1. Conditions

Condition 1: 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 2: 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 2′: 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 1 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. As you can see, this condition does not require y_i to be bounded, so it allows large outliers to occur in the random errors. When the conditional mean of y_i is known to stay in certain range and the related forecasts are relatively restricted, the condition holds. See Subsection 3.1 of Wei & Yang [12] for more discussions on this condition.

Condition 2 generally requires that the estimates of the scale parameters are not too small compared to the truth. Condition 2′ requires that the estimates of the scale parameters are not too far from the truth in both directions.

2.4.2. 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 degree 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 (4). Therefore, $\hat{q}_i^{A_t}$ is the mixture estimator of q_i from the t-AFTER procedure. Let $D(f||g) := \int f \log \frac{f}{g}$ be the Kullback–Leibler divergence between two density functions f and g. Therefore, $E\big(D(q_i||\hat{q}_i^{A_t})\big)$ is a measure of the performances of $\hat{q}_i^{A_t}$ as an estimate of q_i under the Kullback–Leibler divergence at time point i.

Theorem 1. *If the random errors are from a scaled Student's t-distribution with degrees of freedom ν and Condition 2 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 two and Conditions 1 and 2′ 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 2 is satisfied, Theorem 1 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: the

squared relative estimation errors of the scale parameters and the logarithm of the initial weights. This risk bound is obtained without assuming the existence of the 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 two and both Conditions 1 and 2′ are satisfied, Theorem 1 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 parameter estimation. The risk bounds hold even if the the distribution of random errors has tails as heavy as t_3.
3. If there is no prior information to decide the w_j's in (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.

3. The g-AFTER Methodology

In Section 2, the theoretical risk bounds of the combined forecasts from the t-AFTER are provided 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 the 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.

3.1. 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:

$$W_i^{A_g} = \frac{\mathbf{l}_{i-1}^{A_g}}{||\mathbf{l}_{i-1}^{A_g}||_1}, \quad \hat{y}_i^{A_g} = \langle \hat{Y}_i, W_i^{A_g} \rangle, \tag{7}$$

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

$$l_{i-1,j}^{A_g} = l_{i-1,j}^{A_2} + c_1 l_{i-1,j}^{A_1} + c_2 l_{i-1,j}^{A_t}, \tag{8}$$

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 importance 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 random errors are likely to be normally distributed.

3.2. Conditions

Condition 3: 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 one, we have:

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

where s_i is 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 of the estimates of the scale parameters to 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 Subsections 2.2 and 2.3.

Condition 4: When the random 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. 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 (5) and $\mathbf{l}_{i-1}^{A_g}$ is defined in (8). Therefore, $\hat{W}_{i,j}^{A_2}$, $\hat{W}_{i,j}^{A_1}$ 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 (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 2. *If Conditions 3 and 4 hold, then for $\hat{y}_i^{A_g}$ from a g-AFTER procedure, we have:*

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

where:

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

If Condition 1 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 the parameters in Condition 4.

Remarks:

1. Theorem 2 provides a risk bound for more general situations compared to Theorem 1. 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 $\hat{w}_{i,j,k}^{A_t}\ge 0$ for all j and k. Without assuming Condition 1 is satisfied, it follows for any $n\ge 1$:

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

where $w_{j,k}^{A_t}$ is defined the same as that in Subsection 2.3 and:

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

If Condition 1 is also satisfied, then it follows:

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

where C, B_1, B_2 and B_3 are the same as in Theorem 2.

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 random 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 have 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 $\Omega = \{1, 3\}$ as the set of candidate degrees of freedom for the scaled Student's t-distributions considered in the t-AFTER method. 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 a 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 evidence to identify the distribution of the error terms from the three candidate distribution families, we give equal initial weights to the candidate distributions. Therefore, $c_1 = 1$, $c_2 = 2$, $w_j^{A_1} = w_j^{A_2} = 1/J$ 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}.

4.1. Linear Regression Models

4.1.1. 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 one, and the off-diagonal elements are 0.8. The forecasters are generated after the 90th observation, and the combination is generated after the fifth forecast. 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)) of each combination method, where $\hat{y}_i$ is the forecast of y_i from this method. Although we focus on the ASEE in our presentation of the numerical results, another measure, the averaged absolute estimation error (AAEE), $\frac{1}{20}\sum_{i=106}^{125}|m_i - \hat{y}_i|$, is also considered. The main results are similar under the two performance measures. In Sub-subsection 4.2.3, some results are given under both ASEE and AAEE to demonstrate that the comparison results are robust to the selection of the performance measure. Note that, since this is a simulation study, the combined forecasts are compared to 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 (or AAEE) over the 200 datasets is recorded.

We sample β 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 are recorded. The summaries (means and their standard errors) of the 200 sets of ratios are presented.

4.1.2. Results

Three sets of results that correspond to three choices of the number of variables in the true models, *i.e.*, $p_0 = 3, 5, 10$, respectively, are presented in Table 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. It is expected that the t-AFTER and g-AFTER will outperform the L_1-AFTER and L_2-AFTER when forecasting data generating processes (DGPs) with heavy-tailed distributions in the errors. Thus, we run simulations with errors following scaled t_3, t_{10}, double-exponential (DE) and normal distributions. As one can see in Table 1, the t- and g-AFTER are the best forecasters for DGPs with errors coming from t_3, t_{10} and DE distributions. In those cases, the L_1-AFTER also outperformed the L_2-AFTER. In addition, the g-AFTER and the L_2-AFTER are the best forecasters for the normal case. In summary, the t-AFTER and g-AFTER are better choices for heavy-tailed distributions, and the general forecaster g-AFTER performs also very well for DE and normal errors.

Table 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 (0.009)	1.043 (0.003)	1.116 (0.004)	1.028 (0.001)	0.983 (0.003)	0.958 (0.001)	0.926 (0.002)	0.931 (0.001)
At	0.943 (0.002)	0.980 (0.001)	0.983 (0.001)	0.995 (0.001)	0.941 (0.003)	0.955 (0.001)	0.932 (0.001)	0.942 (0.001)
Ag	0.944 (0.002)	0.967 (0.001)	0.974 (0.001)	0.977 (0.001)	0.940 (0.001)	0.950 (0.001)	0.926 (0.001)	0.938 (0.001)

Table 1. *Cont.*

	t_3		DE		t_{10}		$Normal$	
	$\sigma^2=1$	$\sigma^2=9$	$\sigma^2=1$	$\sigma^2=9$	$\sigma^2=1$	$\sigma^2=9$	$\sigma^2=1$	$\sigma^2=9$
				$p_0=5$				
$A2$	1.257 (0.008)	1.066 (0.004)	1.088 (0.003)	1.026 (0.001)	0.980 (0.002)	0.955 (0.001)	0.937 (0.002)	0.927 (0.001)
At	0.950 (0.002)	0.967 (0.001)	0.976 (0.001)	0.982 (0.001)	0.951 (0.001)	0.950 (0.001)	0.943 (0.001)	0.938 (0.001)
Ag	0.951 (0.001)	0.958 (0.001)	0.971 (0.001)	0.970 (0.001)	0.949 (0.001)	0.944 (0.001)	0.939 (0.001)	0.933 (0.001)
				$p_0=10$				
$A2$	1.166 (0.006)	1.056 (0.003)	1.035 (0.002)	0.998 (0.001)	0.968 (0.002)	0.949 (0.001)	0.946 (0.001)	0.929 (0.001)
At	0.950 (0.002)	0.957 (0.001)	0.964 (0.001)	0.965 (0.001)	0.949 (0.001)	0.946 (0.001)	0.948 (0.001)	0.939 (0.001)
Ag	0.945 (0.001)	0.949 (0.001)	0.961 (0.001)	0.955 (0.001)	0.944 (0.001)	0.939 (0.001)	0.942 (0.001)	0.933 (0.001)

Note: The first row shows the distributions of the random errors in the true data-generating regression model, and DE stands for double-exponential distribution. The second row describes the noise variance in data generation. The $A2$, At and Ag stand for the L_2-, t- and g-adaptive forecasting through exponential re-weighting (AFTER) methods, respectively. The parameter p_0 is the number of explanatory variables in the data-generating model. The true parameter (β) values are randomly generated from a uniform distribution, and the candidate forecasts are obtained from linear regressions with 1, 2 and up to the maximum number of explanatory variables. For each set of true parameters, 200 replicated datasets are generated to simulate the mean average squared estimation error (ASEE) for each combination method. The ratio of the mean ASEE of each method over that of the L_1-AFTER is used to measure the relative performance of the competitors. The process is replicated 200 times, each time with independently-generated true β values. The means and their standard errors of the 200 sets of ratios are summarized in this table (the numbers in the parentheses are the standard errors).

4.2. AR Models

4.2.1. Simulation Settings

Let the true model be a $AR(p_0)$ process with random errors from certain distributions and the candidate forecasters be based on $AR(1), AR(2), \cdots, AR(p)$ ($1 \le p_0 \le p$), respectively. For results on asymptotically-optimal model selection for AR models, see, e.g., Ing [24] and Ing *et al.* [25]. 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 from $[-1, 1]$. The β leading to a non-stationary AR model is not considered. Given a valid β, 200 samples with size $n = 125$ from the true model are generated. On each data sample, the candidate forecasters are generated after the 90th observation, and the ASEE of the last 20 forecasts is recorded. Furthermore, the combined forecasts are compared to the conditional means instead of the observations. For each β, the mean ASEE of each combining method over the 200

samples is recorded, and the ratios of the mean ASEEs of the other methods over that of the L_1-AFTER are recorded.

We replicate the generation of p_0's (and β's) 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).

4.2.2. Other Combination Methods

Some other popular combination methods are included in this part and compared to the newly-proposed methods. The 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 [1]) we use in this paper is the version in Hansen [26]. Furthermore, 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 [27], for more details. Hereafter, for example, $BG_{0.9}$ 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 one (without the intercept in the regressions).

4.2.3. Results

In order to demonstrate the advantageous performances of the *t*- and *g*-AFTER for heavy-tailed DGPs in various scenarios, we simulated two major cases for comparison. Tables 2 and 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. See Sub-subsection 4.1.2 for the descriptions of these methods. The other entries are defined as in Table 1. Table 2 presents the results for the cases that the random errors are not (or only mildly) heavy tailed, while Table 3 contains the results when the random errors have significant heavy tails.

One can see that the t- and g-AFTERs consistently outperform all other non-AFTER-based combination methods in all of the simulated situations (heavy tailed or not) and outperform the L_1- and L_2-AFTERs when the random errors have tails heavier than normal. The CLR is competitive because the constraints in its processes make the combination weights of the candidate forecasts relatively more stable and resistant to dramatic changes. The CLR gets more competitive when the random errors have heavier tails. The SA and TM are vulnerable to outliers, which hurts their overall performances. Their ASEEs are over 35% to 150% more than those of the proposed methods. We can see this from both tables.

Table 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)

Note: The first column lists the competing forecast combination methods. The true models for this study are autoregressive (AR) models with the true order randomly generated up to $p = 5$. The true parameters in the AR model are uniformly generated (with the parameters leading to non-stationary AR models removed). The 5 candidate forecasts are obtained from $AR(1)$, $AR(2)$, up to $AR(5)$ models. All other aspects are similar to Table 1. Some other popular combination methods are included in the comparison. The SA, MD and TM methods use the average, median and trimmed mean (removing the largest and smallest before averaging) as the combined forecasts, respectively. The BG method uses the inverse of the historical mean squared forecast errors of the candidate forecasts to assign combination weights. A modified BG method is used with a discount factor $0 < \rho < 1$ to discount the contribution of forecast errors at an earlier time when estimating the variances of the candidates (e.g., Stock & Watson [27]). Here, for instance, $BG_{0.9}$ denotes this BG method with $\rho = 0.9$. The LR method uses linear regression with the actual value as the response and the candidate forecasts as the regressors in linear regression to assign the combination weights. The CLR method is LR with the constraint that the coefficients are non-negative and sum to 1.

Table 3. Simulation results on the AR models with $p = 5$ (heavy tailed) under squared estimation error.

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

Note: In the columns of "log-normal", the σ's are the scale parameters instead of the standard deviations of the log-normal distributions. The setting is basically the same as that in Table 2, only the innovation errors in the true models have heavier tails.

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 and consistent when the tails of the distributions of the random errors get heavier. In both Tables 2 and 3, the CLR is the most competitive method outside the AFTER family, but it still has 3% to 7% larger ASEEs than the new methods on average.

In our settings, similar to many real application situations, using the conditional variances only to assign relative combining weights may not be enough, since some of the candidate forecasters are highly

correlated. This explains why the BG and the discounted BG's are not quite competitive, as seen in Tables 2 and 3. The BG related methods have at least 20% larger ASEEs than the AFTER-based methods on average.

To demonstrate that our results are not sensitive to the performance measure, we redo Table 3 under the AAEE (instead of ASEE), and the comparisons are given in Table 4. The results are similar.

Table 4. Simulation results on the AR models with $p = 5$ (heavy tailed) under absolute estimation error.

	t_3			Log-Normal		
	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 9$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1$
$A2$	1.018 (0.003)	1.019 (0.002)	1.019 (0.008)	0.981 (0.002)	0.997 (0.002)	1.017 (0.003)
At	0.990 (0.002)	0.988 (0.002)	0.993 (0.002)	0.982 (0.002)	0.976 (0.002)	0.975 (0.003)
Ag	0.988 (0.002)	0.986 (0.002)	0.991 (0.002)	0.979 (0.002)	0.978 (0.002)	0.977 (0.002)
SA	1.469 (0.064)	1.666 (0.076)	1.724 (0.080)	1.435 (0.069)	1.543 (0.064)	1.483 (0.064)
MD	1.209 (0.043)	1.314 (0.068)	1.412 (0.094)	1.129 (0.035)	1.279 (0.060)	1.196 (0.037)
TM	1.226 (0.040)	1.367 (0.056)	1.312 (0.085)	1.183 (0.033)	1.331 (0.050)	1.272 (0.040)
BG	1.187 (0.023)	1.272 (0.029)	1.489 (0.035)	1.159 (0.021)	1.245 (0.027)	1.210 (0.023)
$BG_{0.95}$	1.184 (0.022)	1.269 (0.029)	1.401 (0.034)	1.157 (0.021)	1.242 (0.027)	1.206 (0.023)
$BG_{0.9}$	1.181 (0.022)	1.266 (0.029)	1.378 (0.033)	1.156 (0.021)	1.240 (0.027)	1.201 (0.022)
$BG_{0.8}$	1.176 (0.021)	1.260 (0.028)	1.450 (0.033)	1.156 (0.021)	1.237 (0.027)	1.192 (0.021)
$BG_{0.7}$	1.173 (0.021)	1.256 (0.028)	1.352 (0.032)	1.159 (0.021)	1.236 (0.026)	1.185 (0.020)
LR	2.891 (0.084)	2.862 (0.097)	3.647 (1.393)	2.690 (0.074)	2.610 (0.077)	3.296 (0.121)
CLR	1.029 (0.006)	1.025 (0.006)	1.022 (0.006)	1.019 (0.008)	1.004 (0.008)	1.015 (0.006)

Note: The settings are the same as those for Table 3, but this table uses averaged absolute estimation errors (AAEE) instead of averaged squared estimation errors (ASEE) to measure the performance of forecasters. See Sub-subsection 4.1.1 for the detailed definitions of the AAEE and ASEE. The results are similar to Table 3.

5. Real Data Example

The M3-competition data are popular and often used to compare and validate the performances of prediction methods. It contains 3003 micro, industry, macro, financial, demographic and other variables (see [28] for more details). There are 24 different forecast sequences from 24 different candidate forecast models/methods/procedures for each of the 3003 variables (N1 to N3003). For each variable, the last few (6, 8 or 18) observations are not used to train the predictive models, and they are used to evaluate the model performance. Notice that the forecasts are generated all at once (1-, 2-, $\cdots$ and up to 6, 8 or 18 steps ahead) by each forecast model. We use the 1428 variables (N1402 to N2829) with 18 observations as performance evaluating sets to conduct our study because some combination methods need a few forecasts to train the parameters before achieving a reasonable level of reliability.

5.1. Data and Settings

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 $\frac{1}{n_1-n_0+1}\sum_{i=n_0}^{n_1}(y_i - \hat{y}_i)^2$. We use the mean squared forecast errors to measure the prediction performances of the combination methods on each of the 1428 variables. For each variable, the MSFE of each of the other combination methods over the MSFE of the SA is reported. In addition, mean absolute percentage error (MAPE) is also considered.

Specifically, using the same notations as those in Subsection 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. See Sub-subsection 4.1.2 for the descriptions of these methods. The main reason that we use the SA as the benchmark on this real dataset 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 Subsection 4.2 are not considered here.

For each of the variables with 18 forecast periods, the combination starts after the sixth forecasts, and the MSFE of the last nine 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 first and third quartiles (denoted as Q_1 and Q_3, respectively) and the maximum of the 1428 ratios of each method are reported in Table 5. Note that the table also contains the comparisons under the MAPE (all of the other aspects are the same).

Furthermore, the comparison on a subset of M3-competition data is provided. On this subset, the variables are considered to have high potentials to be heavy tailed. All 1428 variables have monthly time intervals. For each of the 1428 variables, there are some training data (about 70 to 128 months). We modeled the training data to find the ones with high potential to have heavy-tailed errors. Specifically, let y_t be the observed value of a variable at time t, and we fit each variable with a model as: $y_t = \beta_0 + \sum_{j=1}^{11}\beta_j I(m_t = j) + \beta_{12}y_{t-1} + \cdots + \beta_{16}y_{t-5}$, where m_t is the month at time point t. We used AIC in backward selection, and the variables with a kurtosis of the forecast errors larger than three were considered to have heavy tails. There are 199 out of 1428 variables that were selected.

Table 5. Results on the 1428 variables of the M3-competition data.

	Mean	**Se**	**Median**	**Min**	Q_1	Q_3	**Max**
$A1$	0.708	0.016	0.649	0.001	0.307	0.994	11.50
	0.758	0.009	0.773	0.038	0.507	0.990	2.901
$A2$	0.697	0.017	0.639	0.001	0.309	0.979	13.32
	0.766	0.010	0.766	0.030	0.517	0.992	4.138
At	0.708	0.015	0.646	0.001	0.312	1.003	8.632
	0.760	0.009	0.769	0.034	0.509	0.993	3.717
Ag	0.696	0.014	0.645	0.001	0.308	0.987	7.710
	0.757	0.009	0.770	0.033	0.508	0.990	3.298
MD	1.050	0.010	1.022	0.002	0.910	1.143	5.341
	1.015	0.005	1.015	0.065	0.944	1.078	2.821
TM	0.990	0.004	1.000	0.002	0.974	1.023	2.437
	0.992	0.002	0.999	0.062	0.984	1.013	1.747
BG	0.784	0.010	0.838	0.001	0.596	0.973	5.227
	0.849	0.006	0.902	0.039	0.758	0.983	3.051
$BG_{0.95}$	0.775	0.010	0.832	0.001	0.582	0.969	7.715
	0.842	0.006	0.896	0.037	0.749	0.981	2.841
$BG_{0.9}$	0.768	0.012	0.825	0.001	0.564	0.966	11.45
	0.835	0.006	0.893	0.036	0.739	0.978	2.643
$BG_{0.8}$	0.758	0.019	0.806	0.001	0.529	0.960	24.08
	0.822	0.006	0.883	0.040	0.709	0.974	2.712
$BG_{0.7}$	0.757	0.031	0.793	0.001	0.503	0.956	43.19
	0.810	0.007	0.870	0.036	0.684	0.971	3.517

Note: For each of the 1428 variables, the methods in the first column are used to combine the 24 forecasts of the last 9 periods of the 18 data points. The mean squared forecast error (MSFE) and mean absolute percentage error (MAPE) of each method are recorded. The ratios of the MSFEs and MAPEs of these methods over that of the SA, respectively, are used as the relative performances of the competing methods. For each forecast combination method other than SA, the mean, minimum, maximum, first quartile, third quartile and the associated standard error of the mean of these ratios based on the 1428 series are summarized in this table. In the table, the summaries of the results based on MSFE are on top of those based on MAPE for each method.

For the heavy-tailed subset, we want to focus on the comparison between the g-AFTER and the non-AFTER methods, because the comparison inside the AFTER family is well addressed in the simulation settings. The reason we choose the g-AFTER instead of the t-AFTER for further comparison is because g-AFTER is practically more efficient, since it performs well even if the signal of heavy tails is not extremely strong. Therefore, for this subset, the benchmark method is the g-AFTER, and the results are reported in Table 6. The comparisons under both the MAPE and the MSFE are provided in the table.

Table 6. Results on the heavy-tailed subset.

	Mean	**Se**	**Median**	**Min**	Q_1	Q_3	**Max**
SA	7.738	1.695	2.259	0.131	1.311	5.244	82.734
	2.044	0.166	1.422	0.327	1.056	2.147	25.784
MD	8.088	2.005	1.912	0.222	1.162	4.974	120.428
	1.998	0.153	1.406	0.477	1.030	2.055	21.229
TM	7.607	1.664	2.299	0.129	1.267	5.175	78.481
	2.014	0.165	1.416	0.316	1.035	2.150	26.039
BG	2.073	0.245	1.266	0.245	0.961	2.160	40.137
	1.349	0.053	1.157	0.468	0.971	1.565	7.845
$BG_{0.95}$	2.017	0.217	1.431	0.241	0.965	2.472	12.551
	1.322	0.048	1.154	0.465	0.965	1.525	6.703
$BG_{0.9}$	1.846	0.182	1.337	0.208	0.958	2.444	10.383
	1.295	0.043	1.114	0.461	0.954	1.497	5.655
$BG_{0.8}$	1.656	0.150	1.340	0.179	0.851	2.074	8.577
	1.246	0.036	1.100	0.454	0.940	1.448	3.985
$BG_{0.7}$	1.536	0.141	1.256	0.158	0.813	1.673	7.746
	1.202	0.032	1.089	0.431	0.928	1.371	3.461

Note: Out of the 1428 variables, 199 of them are identified to have heavy tails in the forecast errors. For these 199 variables, the g-AFTER method is used as the benchmark method for comparison, and all other aspects in the setting are the same as Table 5. Note that for this heavy-tailed subset, we want to focus on the comparison between the g-AFTER and the non-AFTER methods, because the comparison inside the AFTER family is well addressed in the simulation settings. The reason we choose the g-AFTER instead of the t-AFTER for further comparison is because g-AFTER is practically more efficient, since it performs well even if the signal of heavy tails is not extremely strong.

5.2. Results

As one can see from Table 5, the overall performances of the AFTER-based methods are better than the other popular combination methods considered by providing at least 6% to 7% smaller MSFEs on average. It also 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, as seen from the last column.

In the scenario that the the DGPs are believed to have heavy-tailed distributions, the g-AFTER is significantly better than the non-AFTER methods by providing significantly smaller MSFE (about 33% smaller than the best of the competitors on average), as seen in Table 6. Therefore, the robustness of g-AFTER is supported by the M3-competition data.

Table 5 also shows that the AFTERs can occasionally be significantly worse than the SA and other methods. From Table 5, 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 cannot 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.

From Tables 5 and 6, it is clear that the advantages of the t- and g-AFTER hold under both MSFE and MAPE.

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 paper, we propose combination methods designed for cases when forecast errors exhibit heavy-tailed 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 is proposed to solve the estimation problem, and the resulting t-AFTER works well, as seen in the 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 is considered, and the combination process automatically decides which ones are more reasonable by giving them high weights. The numerical results suggest that 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.

In the present numerical work, the numbers of candidate forecasts considered are relatively small. In some situations, there are large numbers of candidate forecasts to begin with. It has been shown in the literature that a proper screening before combining can be beneficial, and information criteria can be used to choose top performers to be combined (see, e.g., Yuan & Yang [29] and Zhang *et al.* [23]). Alternatively, one may also use model confidence sets (see Hansen *et al.* [30] and Ferrari & Yang [31]) to narrow the pool of candidates before applying a combining method. Samuwels & Sekkel [32] provide an interesting comparative study on the effect of screening via a model confidence set of Hansen *et al.* [30], which shows that removing poor candidates indeed improves the final performance of the combined forecast. In the future, it will be useful to investigate how the t- and g-AFTER methods behave when a screening step is applied before combining.

Acknowledgments

We thank the associate editor and two reviewers for their very helpful comments and suggestions for improving the paper. This work was partially supported by the U.S. National Science Foundation Grant DMS-1106576. We thank the Minnesota Supercomputing Institute for providing computing resources.

Author Contributions

All the authors contributed to the formulation of the problem, its solution, numerical work and the writing of the paper.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

A.1

In this subsection, some simple facts are given. They are used in Subsection A.2 of the Appendix.

- 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 arithmetic, 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 $s \ge s_0 > 0$. Use Fact 2 to show that $-\log(s) = \log(1+(1-s)/s) \le (1-s)/s$.

A.2

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:

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

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

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

- Using Fact 2 in Subsection A.1, it follows: $\frac{1}{2}\log\frac{\nu'}{\nu} \leq \frac{1}{2}\frac{\nu'-\nu}{\nu} \leq \frac{1}{2}\frac{|\nu'-\nu|}{\nu}$.
- It is easy to show that:

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

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

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

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

The proof can be completed by combining these steps. □

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:*

$$\int h(x)\log\frac{h(x)}{\frac{1}{s}h\left(\frac{x-t}{s}\right)} \leq C_4(1-s)^2 + C_5t^2,$$

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:

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

□

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,

$$E \log \frac{h(Y)}{\frac{1}{s}h\left(\frac{Y-t}{s}\right)} dy = \log(s) + \frac{1+t^2-s^2}{2s^2} = \frac{1}{2s^2}t^2 + \log(s) + \frac{1-s^2}{2s^2} \leq \frac{1}{2s^2}t^2 + (s-1) + \frac{1-s^2}{2s^2}$$
$$= \frac{1}{2s^2}t^2 + \frac{2s+1}{2s^2}(s-1)^2 \leq \frac{1}{2s_0^2}t^2 + \frac{2s_0+1}{2s_0^2}(s-1)^2.$$

□

A.3

In this subsection, we prove Theorem 1.

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. Therefore, q^n is the estimator of f^n.

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

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

Conditional on all of the information before time point i,

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

By Lemma 1 in Subsection 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 $B_1 = \nu \frac{2+s_0}{2s_0}$. Therefore,

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

From Theorem 1 of Yang [8], there exists a constant C depending on the parameters in Conditions 1 and 2′, 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 2′ instead of Condition 2.

A.4

The essential part of the proof of Theorem 2 is provided in this subsection. We only provide the steps of the proof when the random errors are scaled Student's t-distributed, since the proofs 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 degree of freedom. If Condition 4 holds, then obviously:

$$q^n \geq \sum_{k=1}^{K}\sum_{j=1}^{J}c_2w_{j,k}^{A_t}/G\prod_{i=i_0+1}^{i_0+n}\frac{1}{\hat{s}_{i,j,k}}h_{\nu_l}\left(\frac{y_i-\hat{y}_{i,j}}{\hat{s}_{i,j,k}}\right).$$

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

$$\log\frac{f^n}{q^n} \leq \log\frac{\prod_{i=i_0+1}^{i_0+n}\frac{1}{s_i}h\left(\frac{y_i-m_i}{s_i}\right)}{c_2w_{j^*,k^*}^{A_t}/G\prod_{i=i_0+1}^{i_0+n}\frac{1}{\hat{s}_{i,j^*,k^*}}h_{\nu_{k^*}}\left(\frac{y_i-\hat{y}_{i,j^*}}{\hat{s}_{i,j^*,k^*}}\right)} = \log\frac{G}{c_2w_{j^*,k^*}^{A_t}}+\sum_{i=i_0+1}^{i_0+n}\log\frac{\frac{1}{s_i}h\left(\frac{y_i-m_i}{s_i}\right)}{\frac{1}{\hat{s}_{i,j^*,k^*}}h\left(\frac{y_i-\hat{y}_{i,j^*}}{\hat{s}_{i,j^*,k^*}}\right)}.$$

Similarly, by 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 1.

References

1. Bates, J.M.; Granger, C.W.J. The combination of forecasts. *OR* **1969**, *20*, 451–468.
2. Clemen, R.T. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* **1989**, *5*, 559–583.
3. Newbold, P.; Harvey, D.I. Forecast combination and encompassing. In *A Companion to Economic Forecasting*; Clements, M.P., Hendry, D.F., Eds.; WILEY: Malden, MA, USA, 2002; pp. 268–283.
4. Timmermann, A. Forecast combinations. In *Handbook of Economic Forecasting*; NORTH-HOLLAND: Amsterdam, The Netherlands, 2006; Volume 1, pp. 135–196.

5. Lahiri, K.; Peng, H.; Zhao, Y. Machine Learning and Forecast Combination in Incomplete Panels. 2013. Available online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2359523 (accessed on 10 October 2015).
6. Armstrong, J.S.; Green, K.C.; Graefe, A. Golden rule of forecasting: Be conservative. *J. Bus. Res.* **2015**, *68*, 1717–1731.
7. Green, K.C.; Armstrong, J.S. Simple *versus* complex forecasting: The evidence. *J. Bus. Res.* **2015**, *68*, 1678–1685.
8. Yang, Y. Combining forecasting procedures: Some theoretical results. *Econom. Theory* **2004**, *20*, 176–222.
9. Marinelli, C.; Rachev, S.; Roll, R. Subordinated exchange rate models: Evidence for heavy tailed distributions and long-range dependence. *Math. Comput. Model.* **2001**, *34*, 955–1001.
10. Harvey, A.C. *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economical Time Series*; Cambridge University Press: New York, NY, USA, 2013; p. 69.
11. Zou, H.; Yang, Y. Combining time series models for forecasting. *Int. J. Forecast.* **2004**, *20*, 69–84.
12. Wei, X.; Yang, Y. Robust forecast combinations. *J. Econom.* **2012**, *166*, 224–236.
13. Fernandez, C.; Steel, M.F.J. Multivariate Student-*t* regression models: Pitfalls and inference. *Biometrika* **1999**, *86*, 153–167.
14. Fonseca, T.C.O.; Ferreira, M.A.R.; Migon, H.S. Objective bayesian analysis for the Student-*t* regression model. *Biometrika* **2008**, *95*, 325–333.
15. Kan, R.; Zhou, G. *Modeling Non-Normality Using Multivariate T: Implications for Asset Pricing*; Technical Report; Rotman School of Management, University of Toronto: Toronto, ON, Canada, 2003.
16. Granger, C.W.J.; Ramanathan, R. Improved methods of forecasting. *J. Forecast.* **1984**, *3*, 197–204.
17. Sancetta, A. Recursive forecast combination for dependent heterogeneous data. *Econom. Theory* **2010**, *26*, 598–631.
18. Cheng, G.; Yang, Y. Forecast combination with outlier protection. *Int. J. Forecast.* **2015**, *31*, 223–237.
19. Makridakis, S.; Hibon, M. The M3-Competition: Results, conclusions and implications. *Int. J. Forecast.* **2000**, *16*, 451–476.
20. Inoue, A.; Kilian, L. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *J. Am. Stat. Assoc.* **2008**, *103*, 511–522.
21. Sanchez, I. Adaptive combination of forecasts with application to wind energy. *Int. J. Forecast.* **2008**, *24*, 679–693.
22. Altavilla, C.; de Grauwe, P. Forecasting and combining competing models of exchange rate determination. *Appl. Econ.* **2010**, *42*, 3455–3480.
23. Zhang, X.; Lu, Z.; Zou, G. Adaptively combined forecasting for discrete response time series. *J. Econom.* **2013**, *176*, 80–91.
24. Ing, C.K. Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Stat.* **2007**, *35*, 1238–1277.
25. Ing, C.K.; Sin, C.-Y.; Yu, S.-H. Model selection for integrated autoregressive processes of infinite order. *J. Multivar. Anal.* **2012**, *106*, 57–71.

26. Hansen, B.E. Least squares forecast averaging. *J. Econom.* **2008**, *146*, 342–350.
27. Stock, J.H.; Watson, M.W. Forecasting with many predictors. In *Handbook of Economic Forecasting*; NORTH-HOLLAND: Amsterdam, The Netherlands, 2006; Volume 1, pp. 515–554.
28. M3-Competition data. Available online: http://forecasters.org/resources/time-series-data/m3-comp-etition/ (accessed on 9 October 2015).
29. Yuan, Z.; Yang, Y. Combining Linear Regression Models: When and How? *J. Am. Stat. Assoc.* **2005**, *100*, 1202–1204.
30. Hansen, P.; Lunde, A.; Nason, J. The model confidence set. *Econometrica* **2011**, *79*, 453–497.
31. Ferrari, D.; Yang, Y. Confidence sets for model selection by F-testing. *Stat. Sin.* **2015**, *25*, 1637–1658.
32. Samuels, J.D.; Sekkel, R.M. Forecasting with Many Models: Model Confidence Sets and Forecast Combination; Working Paper; 2013. Available online: http://www.bankofcanada.ca/wp-content/uploads/2013/04/wp2013-11.pdf (accessed on 10 October 2015).