Forecast Combination With Outlier Protection

Gang Cheng[a,*], Yuhong Yang[a,1]

[a] *313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455*

Abstract

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 this paper, we propose a synthetic loss function and apply it on a general adaptive combination scheme for outlier-protective combination of forecasts. Both 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.

Keywords: AFTER, Forecast combination, Outlier protection, Robustness, Loss function, M3-Competition

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/preferences of the forecasters, multiple forecasts are available for the target variable of interest in many applications. In order to get most accurate forecasts by taking advantage of the candidate forecasts, the strategy of forecast combination is often applied.

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. For example, combining via simple averaging (e.g., Stock & Watson, 1999), combining via variance-covariance estimation of the candidate forecasts (e.g., Bates & Granger, 1969), combining via Bayesian model averaging (e.g., Min & Zellner, 1993), combining via regression on candidate forecasts (e.g., Granger & Ramanathan, 1984) and combining via exponential re-weighting (e.g., Yang, 2004) have been studied. Reviews and discussions of the research results are available in Clemen (1989), Newbold & Harvey (2002), Timmermann (2006) and Lahiri et. al (2013).

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

*Corresponding author. Tel: +1 (612) 508-5360

Email addresses: chen2285@umn.edu (Gang Cheng), yangx374@umn.edu (Yuhong Yang)

[1] Tel: +1 (612) 626-8337; fax: +1 (612) 624-8868

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 objective 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 paper, 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 paper is as follows: section 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 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 4. Real data from the M3-Competition (see e.g., Makridakis & Hibon, 2000) are used in section 5 and the results also confirm advantages of our methods. Section 6 concludes the paper. The proofs of the theoretical results are presented in the appendix.

2. Outlier Protective Loss Functions

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.

2.1.1. 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 1 and 2. Specifically, $\hat{Y}_{1,i}$ and $\hat{Y}_{2,i}$ are the forecasts for Y_i made respectively by forecasters 1 and 2 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., forecaster 2 first-order stochastically dominates Forecaster 1), 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.

2.1.2. *Example 1*

In this example, in the same context of section 2.1.1, 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 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). Each 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 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. *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:

$$L_{21}(x|\alpha) = |x| + \alpha\frac{x^2}{m}, \tag{1}$$

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:

$$L_0(e|\gamma_1, \gamma_2) = I(e \geq \gamma_1 \text{ or } e \leq \gamma_2). \tag{2}$$

It can be added to the L_{21}-loss in expression (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

Table 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

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) 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), the **MCP** surrogate $\tilde{L}_0$-loss is:

$$\tilde{L}_0(e|\gamma_1,\gamma_2) = \begin{cases} 1, & \text{if } e \geq \gamma_1 \text{ or } e \leq \gamma_2 \\ 1 - \frac{1}{\gamma_1^2(1-r_1)^2}(e-\gamma_1)^2, & \text{if } r_1\gamma_1 \leq e \leq \gamma_1 \\ 1 - \frac{1}{\gamma_2^2(1-r_2)^2}(e-\gamma_2)^2, & \text{if } \gamma_2 \leq e \leq r_2\gamma_2 \\ 0, & \text{if } \gamma_2 r_2 \leq e \leq \gamma_1 r_1, \end{cases} \tag{3}$$

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 $e = \gamma_2 r_2$.

Fig 1 is an example of $\tilde{L}_0(e|\gamma_1 = 2, \gamma_2 = -2, r_1 = r_2 = 0.75)$.

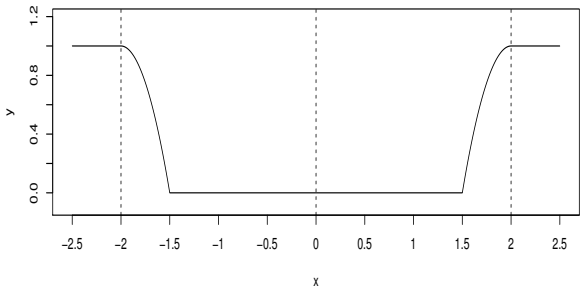

Fig. 1: MCP surrogate of the L_0-loss

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

$$L_{210}(e) := |e| + \alpha_1 \frac{e^2}{m} + \alpha_2 m \tilde{L}_0(e|\gamma_1 m, \gamma_2 m, r_1, r_2), \tag{4}$$

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 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.

2.3. 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.

2.3.1. Scenario 1

Using the scenario in section 2.1.2, the $L_{210}^{(1)}$-loss and the $L_{210}^{(2)}$-loss are defined with common parameters (in expression (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 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.

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

Using the notation from example 1 (section 2.1.2), 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. In Table 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, 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).

3. Forecast Combination with Outlier Protective Loss Function

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:

$$W_{j,t} = \frac{\prod_{k=n_0}^{t-1} \exp\Big(-\lambda L_{210}(y_k - \hat{y}_{j,k})\Big)}{\sum_{j'=1}^{N} \prod_{k=n_0}^{t-1} \exp\Big(-\lambda L_{210}(y_k - \hat{y}_{j',k})\Big)}, \tag{5}$$

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

$$\hat{y}_t^* = \sum_{j=1}^{N} W_{j,t} \hat{y}_{j,t}. \tag{6}$$

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*

Table 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

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*

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

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 Appendix.

3.2. Data Driven L_{210}-AFTER

The choice of the parameter λ in the weighting formula of expression (5) is a difficult issue since it depends on some unknown quantities as discussed in section 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.

3.2.1. $\mathcal{F}_{210}$-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 4.

3.2.2. L_{210}-AFTER with Scale-parameter Estimation

The new weighting formula is:

$$W_{j,t} = \frac{\prod_{k=n_0}^{t-1} \frac{1}{\sqrt{\hat{\delta}_{j,k}}} \exp\left(-L_{210}(y_k - \hat{y}_{j,k})/\hat{\delta}_{j,k}\right)}{\sum_{j'=1}^{N} \prod_{k=n_0}^{t-1} \frac{1}{\sqrt{\hat{\delta}_{j',k}}} \exp\left(-L_{210}(y_k - \hat{y}_{j',k})/\hat{\delta}_{j',k}\right)}, \quad (7)$$

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

The proof of Theorem 2 is provided in the Appendix.

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.

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:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p_0} X_{p_0} + \epsilon, \tag{8}$$

where $1 \le p_0 \le 5$ 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 (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 (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;

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.

4.2. The Competiting 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 (LRC) and the existing AFTER methods.

Specifically, the simple average strategy uses the mean of the forecasts as the combined forecast; the trimmed mean method here removes the largest and the smallest forecasts before averaging; MD uses the median of the candidate forecasts as the combined forecast; and the BG method used here is exactly the same as that in Hansen (2008), which weights each candidate forecasts by the reciprocal of their estimated variances of the error distributions at the time points of combining. The combination via linear regression (LR) uses the candidate forecasters as predictors and the univariate variable to be predicted as response in a linear regression setting. The combination weights are the ordinary least squares estimates. The intercept is included. The combination via constrained linear regression (LRC) is a modification from the LR: it constraints the coefficients (no intercept is considered) to be non-negative with sum 1.

4.3. Scenarios

4.3.1. 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 $1 \le i, j \le 5$. The p_0 in expression (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 Tables 3, 4 and 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 $1 \le i, j \le 5$ are not included because they provide basically the same stories.

Note that, in Table 3, the values above the parentheses are the means of the corresponding ratios or differences relative to the L_1-AFTER. The values in the parentheses are the corresponding medians. For example, the number 0.825 under the row L_2 and column L_2A is the mean of the ratios of the L_2-losses of the combined forecasts from the L_2A over that of the L_1A, while the number 0.799 is the median of the ratios. The number -0.280 under the row L_0 and column L_2A is the mean of the differences between the L_0-losses (number of outliers) of the combined forecasts from the L_2A and the L_1A. 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 4, the columns 2-5 (6-9) are the means of the ratios of the L_2-losses (L_1-losses) of the combined forecasts from the $L_{210}A$ over those from the L_1A.

In Table 5, the columns 2-5 are the mean differences of the number of outliers in the combined forecasts from the $L_{210}A$-AFTER and those from the L_1A. The last 4 columns are the probabilities that the L_{210}-AFTERs provide forecasts with fewer outliers than that of the L_1-AFTER, respectively. The standard errors for all the values are smaller than 10^{-3}.

4.3.2. 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 Tables 6, 7 and 8 and they are organized in the same ways as those for Scenario 3. Note that the standard errors for all the values in Tables 6 and 8 are smaller than 10^{-3}.

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.

4.4.1. 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 (Tables 3-8) , 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 parameters in the L_{210}-AFTER should not be done carelessly. A general guideline of choos-

ing the parameters efficiently is presented in section 4.4.3.

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.

4.4.2. *The L_{210}-AFTER vs. Other Methods*

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

1. Overall, from Tables 3-8, 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.

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

From our investigations in sections 4.4.1 and 4.4.2, 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.

Tables 3-8 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.
2. From Tables 5 and 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.

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

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

	L_2A	LR	LRC	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)

Table 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 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 6: Popular combination methods under the L_2-, L_1- and L_0-losses (Scenario 4)

	L_2A	LR	LRC	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 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 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

methods (see, e.g., Makridakis & Hibon, 2000; Armstrong, 2007).

Among all the 3003 variables, there are 6 consecutive forecasts by each forecaster for the yearly series, 8 forecasts for quarterly series, and 18 forecasts for monthly series. 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, respectively). 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.

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.

5.2. The Procedures

5.2.1. 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 9 (under the symmetric L_0-loss) and Table 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 11 (under the symmetric L_0-loss) and Table 12 (under the asymmetric L_0-loss).

5.2.2. 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 $\alpha_2 \in \{3, 0.15\}$. Hereafter, for example, the $L_{210}A^{0.15,3}$ stands for a L_{210}-loss function with $(\alpha_1, \alpha_2) = (0.15, 3)$.

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.

5.3. Results

5.3.1. Comparing Different Schemes on the 1428 Variables

Tables 9 and 10 provide the the comparison among methods over the 1428 variables under the L_2-, L_1- and L_0-losses.

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.

5.3.2. 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 Tables 11 and 12, which are organized in the same way as Tables 9 and 10.

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.

Table 9: Relative performance over the SA on the M3-competition Data (Symmetric case)

		TM	MD	BG	L_1A	L_2A	$L_{210}A^{.15,3}$	$L_{210}A^{.15,.15}$	$L_{210}A^{.03,3}$	$L_{210}A^{.03,.15}$
L_2	Mean	0.990	1.048	0.783	0.717	0.702	0.887	0.880	0.845	0.853
	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
L_1	Mean	0.992	1.013	0.851	0.770	0.765	0.825	0.823	0.812	0.811
	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 10: Relative performance over the SA on the M3-competition Data (Asymmetric case)

		TM	MD	BG	L_1A	L_2A	$L_{210}A^{.15,3}$	$L_{210}A^{.15,.15}$	$L_{210}A^{.03,3}$	$L_{210}A^{.03,.15}$
L_2	Mean	0.990	1.048	0.783	0.717	0.702	0.886	0.880	0.842	0.853
	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
L_1	Mean	0.992	1.013	0.851	0.770	0.765	0.824	0.822	0.811	0.811
	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.

Table 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	$L_{210}A^{.15,3}$	$L_{210}A^{.15,.15}$	$L_{210}A^{.03,3}$	$L_{210}A^{.03,.15}$
	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 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	$L_{210}A^{.15,3}$	$L_{210}A^{.15,.15}$	$L_{210}A^{.03,3}$	$L_{210}A^{.03,.15}$
	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.

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.

Acknowledgement

We sincerely thank the two reviewers and the AE for their very helpful comments and suggestions for improving our work. We also thank the Minnesota Supercomputing Institute for providing computing resources.

Appendix

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, $\frac{\alpha_2}{\alpha_1} < \min\{\gamma_2^2(1-r_2)^2, \gamma_1^2(1-r_1)^2\}$ 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:

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

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

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

where

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

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 $2\bar{c}\lambda \le t_0$,

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

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

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

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

Thus, we have,

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

Further, similarly in Yang (2004),

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

Since the analysis is based on an arbitrary j, so

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

This completes the proof of Theorem 1.

Proof of Theorem 2.

For $\delta > 0$, recall

$$h(\delta) := \int \exp\Big(-\frac{L_{210}(x)}{\delta}\Big)dx. \tag{9}$$

Since

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

then, from (9) and Condition 3,

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

where

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

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

$$\sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\xi_2 \le h(\delta) \le \sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}. \tag{10}$$

Recall

$$g(x|\delta) := \frac{1}{h(\delta)}\exp\Big(-\frac{L_{210}(x)}{\delta}\Big). \tag{11}$$

Then, as in Yang (2004),

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

where

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

Then

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

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

$$\Big|\log\Big(\frac{h(\hat{\delta}_{j,i})}{h(\delta_i)}\Big)\Big| \le \xi_3|\hat{\delta}_{j,i}-\delta_i| \le A\xi_3\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i} = c_1\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i} \tag{12}$$

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

$$\begin{aligned} &h(\delta_i)E_iL_{210}(y_i-\mu_i) \qquad (13)\\ &= \int \exp(-\frac{1}{\delta_i}L_{210}(x))L_{210}(x)dx \\ &\le \int \exp(-\frac{\alpha_1}{m\delta_i}x^2)\frac{\alpha_1}{m}x^2dx + \int_{\frac{\alpha_1 A}{m}x^2\le 1} L_{210}(x)dx \\ &\Big(\text{For } \frac{\alpha_1 A}{m}x^2 \ge 1, \exp(-\frac{L_{210}(x)}{\delta_i})L_{210}(x) \le \exp(-\frac{\alpha_1 x^2}{m\delta_i})\frac{\alpha_1 x^2}{m}\Big) \\ &= 2\sqrt{\pi}\frac{m}{2\alpha_1}\delta_i^{3/2} + \xi_4\delta_i^{3/2} \\ &= \xi_5\delta_i^{3/2} \qquad (14) \end{aligned}$$

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

Then,

$$E_iL_{210}(y_i-\mu_i) \le \frac{\xi_5\delta_i^{3/2}}{\sqrt{\delta}\sqrt{\frac{m\pi}{\alpha_1}}\xi_2} = \xi_6\delta_i, \tag{15}$$

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:

$$\begin{aligned} \Big|(\frac{1}{\hat{\delta}_{j,i}} - \frac{1}{\delta_i})L_{210}(y_i-\mu_i)\Big| &\le \Big|\frac{1}{\hat{\delta}_{j,i}} - \frac{1}{\delta_i}\Big|\xi_6\delta_i \\ &\le A^2\xi_6\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i} = c_2\frac{|\hat{\delta}_{j,i}-\delta_i|}{\delta_i}, \end{aligned} \tag{16}$$

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

Similarly,

$$\Big|\frac{1}{\hat{\delta}_{j,i}}\Big(L_{210}(y_i-\hat{y}_{j,i}) - L_{210}(y_i-\eta_i)\Big)\Big| \le B\frac{|L_{210}(y_i-\hat{y}_{j,i}) - L_{210}(y_i-\eta_i)|}{\delta_i}. \tag{17}$$

Therefore, from (15), (16) and (17), it is true for any j

that, for more details

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

So,

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

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.

References

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.

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

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

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.

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.

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

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.

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.

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

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.

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.

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