Combining Forecasts for Universally Optimal Performance

Wei Qian[a], Craig A. Rolling[b], Gang Cheng[b], Yuhong Yang[b]

[a] *Department of Applied Economics and Statistics, University of Delaware*
[b] *School of Statistics, University of Minnesota*

Abstract

There are two potential directions of forecast combination: combining for adaptation and combining for improvement. The former direction targets the performance of the best forecaster, while the latter attempts to combine forecasts to improve on the best forecaster. It is often useful to infer which goal is more appropriate so that a suitable combination method may be used. This paper proposes an AI-AFTER approach that can not only determine the appropriate goal of forecast combination but also intelligently combine the forecasts to automatically achieve the proper goal. As a result of this approach, the combined forecasts from AI-AFTER perform well universally in both adaptation and improvement scenarios. The proposed forecasting approach is implemented in our R package `AIafter`, which is available at `https://github.com/weiqian1/AIafter`.

Keywords: AFTER, combining forecasts, model averaging, regression, statistical tests

1. Introduction

In many forecasting problems, the analyst has access to several different forecasts of the same response series. These forecasts might arise from models of known structure to the analyst or they may be generated by mechanisms that

are unknown. To utilize these candidate forecasts to accurately predict future response values, the analyst can either select one of the forecasting procedures that seems to perform well or combine the candidate forecasters in some way.

Much work has been devoted to the merits of forecast combination that may take both frequentist and Bayesian-type approaches for various prediction and forecasting scenarios. Many frequentist methods design and employ performance-based criteria to estimate theoretically optimal weights for linear combination to seek forecasting performance potentially superior to any of the individual candidate forecasts. The early classical work of Bates and Granger (1969) proposed the minimization of mean square error criteria to pursue the optimal weights estimated through forecast error variance matrix, and Granger and Ramanathan (1984) formulated different linear regression frameworks for the optimal weights. Adopting minimization of other performance measures like forecast cross validation and information criteria, weighted averaging strategies have been specifically developed for certain classes of known statistical forecasting and prediction models such as factor models (Cheng and Hansen, 2015), generalized linear models (Zhang et al., 2016), spatial autoregressive models (Zhang and Yu, 2018), among many useful others. With the popularity of performance-based combination methods, it is also well-known that estimated weights from sophisticated methods can often deviate much away from the targeted theoretical optimal weights due to weight estimation error and uncertainty (e.g., Smith and Wallis, 2009; Claeskens et al., 2016); this important and well-studied factor, among other possible factors such as structural break and new information (Lahiri et al., 2017; Qian et al., 2019b), contributes to the phenomenon of forecast combination puzzle (Stock and Watson, 2004a; Hendry and Clements, 2004) that in practice, simple equally-weighted averaging or its variants may outperform sophisticated alternatives.

Different from the combining strategies that directly aim to obtain the optimal weights to potentially improve on all the candidate forecasts, a class of aggregation methods recursively updates the combining weights through online re-weighting schemes (see, e.g., Yang, 2004; Lahiri et al., 2017), and these methods typically take a less ambitious objective and only aim to match the performance of the best candidate forecast, with the promise of having smaller cost for weight estimation. Indeed, Yang (2004) studied a representative of these methods called AFTER and showed non-asymptotic forecast risk bounds that illustrate theoretically the heavier cost in forecast risk from attempting to achieve improved forecasts than only aiming to match the best-performing original candidate forecast; Lahiri et al. (2017) showed asymptotically that AFTER tends to impose all unit weights to the best-performing candidate under some mild conditions. Lahiri et al. (2015) also suggested that the less aggressive combining objective is reasonable in consideration of forecast uncertainty under some appropriate measures. Besides the aforementioned frequentist approaches, Bayesian model averaging methods (Hoeting et al., 1999a; Steel, 2011; Forte et al., 2018) are also in alignment with the objective of adapting to the best original candidate forecast since the data generating process is often assumed to be one of the candidate forecasting models (De Luca et al., 2018), while these models are required to be known. Although different combination methods have been designed with different objectives, as neither information on data generating processes nor candidate forecasts' underlying statistical models are necessarily available in practice, it is usually unknown *a priori* to an analyst whether it is feasible to achieve improved forecast combination performance over all the original candidates, which could lead to improperly choosing combination methods and undesirable forecasting performance; for example, blindly applying the AFTER method is expected to be under-performing if lin-

ear combination of forecasts via optimal weighting schemes much outperforms each original candidate.

Despite ongoing issues on which combining methods to use and how to develop new combining methods for better performance, there has been a general consensus from existing empirical and theoretical research that a combination of available candidate forecasts carries important benefits and often produces better results than the selection of a single forecast candidate (e.g., Stock and Watson, 2003; Yang, 2003; Kourentzes et al., 2019). In a recent open forecast competition named the M4 competition (Makridakis et al., 2020), 100,000 time series were provided as a large-scale testing ground to assess the performance of forecasting methods, and over sixty research teams submitted their own forecasting results based on various methods for principled evaluation; notably, the benefits of forecast combination over selection have been re-confirmed as one of the main conclusions in the highlight results of the M4 competition (Makridakis et al., 2018). The following two reasons can partially contribute to the benefits. First, identifying which forecasting procedure is the best among the candidates often involves substantial uncertainty. Depending on the noise that is realized in the forecast evaluation period, several different candidates may have a good chance of being selected as the best one by a selection procedure, and the winner-takes-all approach of forecast selection often results in post-selection forecasts having high variance. Second, different pieces of predictive information may be available to different forecasters; in this situation, a combination of the candidate forecasts has the potential to outperform even the best original candidate procedure due to the sharing of information from combining the forecasts.

These two benefits of combining forecasts are closely related to the two different objectives of combining methods we briefly discussed earlier. Yang (2004) formally distinguishes these two objectives as combining for adaptation (CFA)

and combining for improvement (CFI), where CFA targets the first benefit and CFI targets the second. Specifically, the objective of combining for adaptation is to achieve the best performance among the original candidate procedures while reducing the variance introduced by selection uncertainty. Combining for improvement, on the other hand, aims to improve on even the best procedure in the original candidate set by searching for an optimal combination of the candidates and directly estimating the optimal weights using performance-based criteria. Either of the two goals of forecast combination may be favored, depending on the nature of the data-generating process and the candidate forecasts. Taking an approach for adaptation when improvement is more appropriate can lead to missed opportunities to share information and improve on the original candidate forecasts; on the other hand, taking an approach for improvement when adaptation is more appropriate may result in elevated forecast risk (defined in Section 2.1) from the heavier cost of the more ambitious CFI objective. More detailed explanation with heuristic illustration on this issue will be given in Section 2.

In this paper, we intend to highlight the connection and relationship between CFA and CFI with a testing procedure to assess whether we can improve on the best individual forecaster and then design a new method to capture the benefits of combining under either goal given the data and candidate forecasts at hand. Specifically, our proposal is based on the AFTER method (Yang, 2004), which is particularly designed for the combining goal of adaptation, since it is guaranteed to perform nearly as well as (rather than significantly improving on) the best candidate forecasting procedure, without knowing in advance which procedure is best. However, what if all the original candidate forecasters perform poorly? In many situations, it is still possible to combine these candidates (sometimes referred to as "weak learners") into a forecast to improve over even

the best original candidate and attain the goal of combining for improvement. Correspondingly, we propose an important extension of the AFTER method to combine for both adaptation (A) and improvement (I) so that it can also adapt to the goal for improvement when such a strategy has evident potential; for brevity, we will call the proposed method AI-AFTER. We also implement the proposed method in a user-friendly R package named `AIafter`, which is available at `https://github.com/weiqian1/AIafter`.

Although this paper is focused on point forecast, it is worth noting that significant progress has been made in literature for the important topics on probability/density forecast combination and interval/quantile forecast combination (Granger et al., 1989; Wallis, 2005). In particular, the density/probability forecast combination methods generally need assumptions on statistical modeling forms for candidate forecasts or data generating processes. For example, Clements and Harvey (2011) established the optimal combination forms under several plausible data generating processes and studied the associated estimation of optimal weights under certain performance-based criteria. Following earlier work of Hall and Mitchell (2007), Geweke and Amisano (2011) considered optimal weighting schemes for linear combination of candidate predictive model densities and showed that in contrast to Bayesian model averaging, their method with optimal weights based on predictive scoring rules intends to achieve performance substantially better than any candidate predictive model. On the other hand, interval/quantile forecast combination methods (Granger et al., 1989) consider time-varying heteroscedastic forecast errors whose non-i.i.d. distributions can be estimated flexibly with different parametric or nonparametric density estimation approaches, and they do not require an analyst to have prior knowledge of either statistical models or their parameters that lead to any of the candidate forecasts. For example, Trapero et al. (2019) proposed to obtain

optimal combination through minimization of quantile loss function criterion, while Shan and Yang (2009) proposed an AFTER-type method for quantile forecast combination that can perform nearly as well as the best original candidate. In this sense, the discussion and proposal of this paper on addressing the two different objectives of CFA and CFI for point forecast combination may be also relevant to the interval/quantile forecast combination, which should deserve separate careful study on its own right. We leave the interesting yet challenging extension beyond point forecast for the interval/quantile forecast combination to future investigation.

The remainder of the paper is structured as follows. In Section 2 we formally define the combining goals of adaptation and improvement, and we present illustrative simulations and examples that favor either of the combining objectives. Section 3 describes a statistical test for the potential to combine for improvement. The AI-AFTER method of combining forecasts for either adaptation or improvement is described in Section 4. Simulation and real data evaluation are given in Section 5 and Section 6, respectively. Section 7 gives brief concluding remarks.

2. Two Objectives of Forecast Combination

The objectives of combining for adaptation/improvement (CFA vs. CFI) are understood in terms of the forecast risks they aim to achieve. We will see that the target risk of combining for improvement (that is, the risk using the optimal weights for combining) is always upper bounded by the target risk of combining for adaptation (that is, the risk of the best original candidate forecast). On the other hand, because of the extra higher cost in forecast risk that can be introduced by combining for improvement, CFA may be more favorable if there is little or nothing to be gained by pursuing improvement over the best original

forecast.

2.1. Definitions

Our goal is forecasting a continuous random variable Y with values $Y_1, Y_2, \ldots$ at times $i = 1, 2, \ldots$, respectively. A general forecasting procedure δ is some mechanism that produces forecasts $\hat{y}_{\delta,i}$ for $i \geq 1$. The procedure δ may be built on a statistical model or on outside information, but here we do not assume the model or any other information about the procedure is known. For the combining problem, we start with a collection of M forecasting procedures $\Delta = \{\delta_1, \delta_2, \ldots, \delta_M\}$, which are the original candidate forecasts.

A forecasting procedure ψ is said to be a combined forecast based on Δ if each $\hat{y}_{\psi,i}$ is a measurable function of $Y_1, \ldots, Y_{i-1}$ and $\hat{y}_{\delta_j,l}$ for $1 \leq l \leq i$ and $1 \leq j \leq M$. Let Ψ be a class of combined forecasting procedures based on Δ. Given any $\psi \in \Psi$, let $R(\psi; n)$ be the forecast risk, where $R(\psi; n) = \sum_{i=1}^{n} \mathrm{E}(Y_i - \hat{y}_{\psi,i})^2$ under quadratic loss. Denote by $\psi^* = \psi^*(\Psi, n) = \mathrm{argmin}_{\psi \in \Psi} R(\psi; n)$ the choice that minimizes the forecasting risk $R(\psi; n)$ over all $\psi \in \Psi$, and let $R(\Psi; n) = R(\psi^*; n)$ denote the minimum risk in the class Ψ.

Definition 1. Let $\Psi_0 = \Delta$ be the collection of original candidate forecasts. A forecast combination method that combines for adaptation (CFA) is one that targets the forecast risk $R(\Psi_0; n)$ and thus aims to perform (nearly) as well as the best original candidate in Ψ_0 in terms of forecast risk.

Definition 2. Let Ψ_L be the collection of all linear combinations of the original candidate forecasts. In other words, any member ψ of Ψ_L produces forecasts of the form

$$\hat{y}_{\psi,i} = \sum_{j=1}^{M} w_{i,j} \hat{y}_{\delta_j,i}, \qquad i \geq 1, \tag{1}$$

where $w_{i,j}$'s are weights (or combining coefficients). A forecast combination method that combines for improvement (CFI) is one that targets the forecast

risk $R(\Psi_L; n)$ and estimates optimal combining weights for forecast risk through minimization of forecast performance-based criteria; thus a method for improvement aims to perform nearly as well as the optimal linear combination of the original candidates in terms of forecast risk. It is also possible to construct further enlarged classes extending from Ψ_0 or Ψ_L; for example, one could consider other linear combining (Wang et al., 2014) or combining the forecasts in a nonlinear fashion.

Most forecast combination methods to date have aimed to combine for improvement, starting with the seminal work of Bates and Granger (1969), which aims to derive optimal weights for the special situation of two forecasts in (1), and the work of Granger and Ramanathan (1984), which aims for more general situations with optimal weights estimation by linear regression. Many subsequent papers have attempted to derive optimal combination weights in (1) (e.g., Hansen, 2008; Hsiao and Wan, 2014; Claeskens et al., 2016, among many others), and these methods can be characterized as targeting $R(\Psi_L; n)$, the minimum risk in Ψ_L under certain conditions.

Methods that combine for adaptation typically do not explicitly use the covariances between forecasters but construct combining weights based on each candidate's merit as a predictor. Bayesian model averaging (BMA; e.g., Hoeting et al., 1999b) is an example of CFA methods in the parametric setting. The method of AFTER (Yang, 2004) is flexible to combine different types of forecasts for adaptation by using each candidate's forecasting performance. In the rest of the paper, we will use AFTER as the representative method designed for CFA; to keep the paper self-contained, we briefly describe the AFTER algorithm here. Specifically, following notations above, suppose there are candidate forecasting procedures $\{\delta_1, \cdots, \delta_M\}$. Then at time point i, the AFTER weight assigned to

δ_j is generated based on its previous relative forecast performance as

$$w_{i,j} = \frac{\left(\prod_{t=1}^{i-1} \hat{\sigma}_{j,t}\right) \exp\left(-\lambda \sum_{t=1}^{i-1} \phi\left(\frac{Y_i - \hat{y}_{\delta_j,t}}{\hat{\sigma}_{j,t}}\right)\right)}{\sum_{j'=1}^{M} \left(\prod_{t=1}^{i-1} \hat{\sigma}_{j',t}\right) \exp\left(-\lambda \sum_{t=1}^{i-1} \phi\left(\frac{Y_i - \hat{y}_{\delta_{j'},t}}{\hat{\sigma}_{j',t}}\right)\right)}, \tag{2}$$

where $\hat{\sigma}_{j,t}$ is set to be the sample standard deviation of previous forecast errors of δ_j prior to the observation of Y_i, and $\phi(\cdot)$ is set to be the squared error loss $\phi(x) = x^2$.

Note that in Definition 1, a forecast combination method is *not* required to have all weights but one being exactly zero; rather, the performance of a forecast combination for adaptation in terms of forecast risk should be almost as good as that of the best candidate in $\Psi_0 = \Delta$. In this sense, the AFTER method described above is indeed designed for CFA: the risk of AFTER forecasts can be explicitly compared to $R(\Psi_0; n)$, which shows that under mild conditions, the risk of AFTER is guaranteed to be upper bounded by $R(\Psi_0; n)$, plus a small additive penalty. Within the CFA family, the strategy is to assign the best candidate(s) the highest weight and therefore approach the performance of the strongest candidate as the information regarding which candidate is the best becomes more reliable (Lahiri et al., 2017).

Clearly, since $R(\Psi_L; n) \leq R(\Psi_0; n)$, combining for improvement intends to achieve a lower risk and is therefore the more ambitious goal than combining for adaptation; nevertheless, $R(\Psi_L; n)$ is also harder to achieve and combining for improvement tends to incur higher extra cost than that of combining for adaptation. Although much earlier work has derived the optimal combination weights in (1) under certain conditions, the difficulty and instability in estimating the weights can lead to sub-optimal empirical performance for methods that pursue the goal of improvement. These issues are well understood as discussed in, e.g., Smith and Wallis (2009) and Claeskens et al. (2016). In addition, when one or more of the original candidate forecasts capture the true underlying data-

generating process well or are otherwise able to provide very accurate forecasts, $R(\Psi_L; n)$ and $R(\Psi_0; n)$ may be very close. In these situations, the cost one pays by searching over the larger class Ψ_L will be greater than the gain from pursuing $R(\Psi_L; n)$ over $R(\Psi_0; n)$. Two simulations are given in Section 2.2 for empirical illustration, followed by two examples in Section 2.3 with more explanation.

2.2. Illustrative simulations

We consider two illustrative simulations to help appreciate the different forecasting performance of the CFA and CFI methods.

Simulation 1. Consider a data generating process with two variables where

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{3}$$

ε_i, X_{i1} and X_{i2} are i.i.d. $N(0, 1)$ with the true parameters $(\beta_0, \beta_1, \beta_2) = (0, b_0, b_0)$. Also consider the following two models to generate the candidate forecasts with their parameters estimated by ordinary least squares, where the forecast (4) adds no additional information about the data generating process for (5):

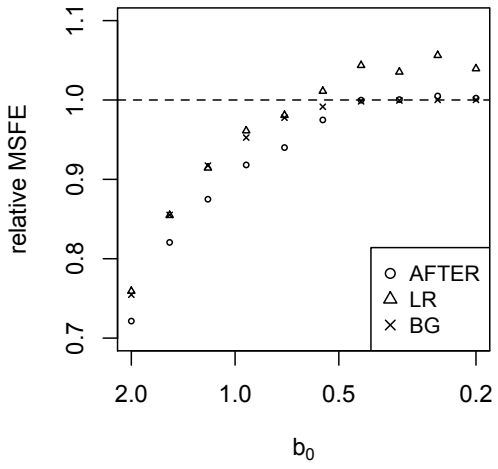
$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{4}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \tag{5}$$

To see that this data scenario should favor combining for adaptation, we perform simulation experiment to evaluate the performance of AFTER, which is designed for CFA; for comparison, we also consider Bates-Granger (BG or BG_1) and linear regression (LR) methods, which are designed for CFI to estimate the optimal weights (Bates and Granger, 1969; Granger and Ramanathan, 1984). The true parameter b_0 for the data generating process (3) takes different values that are evenly spaced between 2.0 and 0.2 for a decreasing sequence of signal strengths. The details on the candidate forecast generation and evaluation are described in Section 5; briefly, with $n_{\text{train}} = 30$ data observations

only for training initial parameters of the candidate forecast models, we use $n = 30$ subsequent observations to generate the candidate forecasts and additional $n_{\text{eval}} = 20$ observations for forecast risk evaluation. Each combining method is evaluated using the mean squared forecast error (MSFE) relative to the MSFE of simple equally weighted averaging (SA) so that the relative MSFE of SA is always at the baseline reference value of 1.0 (to be shown as dashed lines in Figures 1 and 2). The averaged relative MSFEs from repeated experiment are summarized in Figure 1, which shows that AFTER performs favorably and it is appropriate to target the CFA objective. AFTER enjoys the significantly lower forecast risk than the CFI alternatives when the signal b_0 is relatively large and has performance comparable to BG and SA when the signal becomes weak and forecast risk is dominated by the random error variance; on the other hand, LR consistently underperforms, which can be attributed to the large cost from estimation errors and uncertainty in estimating the optimal weights.

Figure 1: Relative forecast performance of CFA vs. CFI methods in Simulation 1.



Simulation 2. Next consider a possible scenario that different candidate forecasts may have different information sets. The data generating process remains

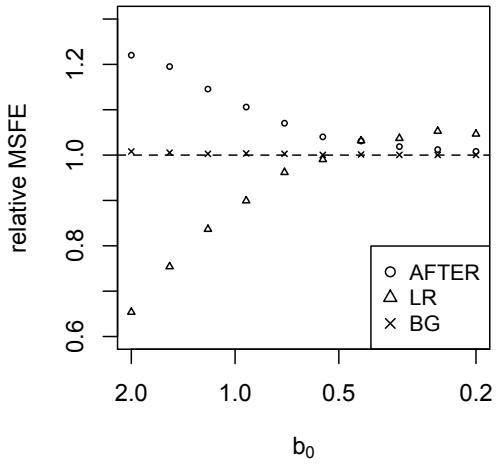
the same as (3) of Simulation 1, but the candidate forecasts are generated by models with different variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i.$$

To see that this data scenario should favor combining for improvement, we apply the same experiment and summarize the averaged relative MSFEs in Figure 2. In stark contrast to Simulation 1, AFTER performs poorly compared to other CFI alternatives when the signal b_0 is relatively large; this is not surprising since AFTER is only designed to perform almost as well as the best candidate forecast, and the optimal linear combination of the two candidate forecasts can much reduce the forecast risk of any single candidate. While LR underperforms again from unstable weight estimation when the signal b_0 becomes weak, AFTER does not seem to perform better than either BG ro SA.

Figure 2: Relative forecast performance of CFA vs. CFI methods in Simulation 2.



2.3. Examples

Extending from the previous empirical illustration, we next give two examples (including nested model scenario and weak learner model scenario) with explanation and discussion on their favored combining objectives.

Example 1 (Nested Models)**.** Competing forecasts are those from models of the same family with different variable subsets and consider a data-generating process where

$$Y_i = \beta_0 + \sum_{j=1}^{p_0} \beta_j X_{ij} + \varepsilon_i, \tag{6}$$

ε_i are i.i.d. with mean 0, and $p \geq p_0$ predictors are available. A common variable selection practice when p is large is to generate a sequence of nested models by doing forward selection. Suppose a forward selection procedure generates the following sequence of models:

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \varepsilon_i \\
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
&\vdots \\
Y_i &= \beta_0 + \sum_{j=1}^{p_0} \beta_j X_{ij} + \varepsilon_i \\
&\vdots \\
Y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i
\end{aligned} \tag{7}$$

When n is large enough to offset the error involved in estimating the forecast model parameters, the forecast generated by (5) above will have risk $R(\Psi_0; n)$. Furthermore, since none of the other models in the sequence add any information about the data generating process to (7), we see that $R(\Psi_0; n) \simeq R(\Psi_L; n)$. In this case, methods that combine for adaptation and those that combine for improvement target the same risk, but it is more difficult for improvement methods to achieve the target risk because of the extra cost from the need to search

over a larger class of combined procedures and the associated estimation errors/uncertainty from optimal weights (Yang, 2004). This example gives the scenario that should favor combining for adaptation.

Example 2 (Weak Learners)**.** Consider a data-generating process where

$$Y_i = \beta_0 + \sum_{j \in \Omega} \beta_j X_{ij} + \varepsilon_i, \tag{8}$$

ε_i are i.i.d. with mean 0, p predictors are available, and the true active set Ω is some subset of $\{1, \ldots, p\}$. Now suppose the candidate forecasting procedures have the form:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \\ Y_i &= \beta_0 + \beta_2 X_{i2} + \varepsilon_i, \\ &\vdots \\ Y_i &= \beta_0 + \beta_p X_{ip} + \varepsilon_i. \end{aligned}$$

This situation could happen if, for example, the p forecasting procedures represent expert forecasts with different information sets. The X_{ip} could be thought of as the information available to expert p at time i. In this scenario, an appropriate combination of the procedures could potentially perform much better than the best individual procedure; that is, $R(\Psi_L; n)$ may be much less than $R(\Psi_0; n)$. The classical methods, which attempt to combine for improvement and aim to improve on the best individual candidate, could become favorable combining choices in this scenario.

In the preceding examples, it seems clear whether CFA or CFI should be the favorable goal given the knowledge on both the data generating process and the candidate forecasting models. However, without such knowledge in practice, it becomes difficult to determine the right goal of combination. The information or model underlying each of the candidate forecasting procedures may be

proprietary or otherwise unknown to the analyst. Additionally, the analyst's set of candidate forecasts may be a mix of statistical models, machine learning algorithms, and expert opinions. In such situations, it is not clear whether one should combine for adaptation or improvement. For example, blindly applying AFTER in Example 2 (and Simulation 2) without knowing that the underlying scenario favors the objective of improvement could result in under-performing forecast outcomes. To address the resulting dilemma of not knowing underlying data scenario's favorable combining objective and how to choose a proper forecast combination method, in the next section, we propose a statistical test that measures evidence of the potential to combine for improvement. The information provided by this test will allow further understanding of the combination problem and lead to the forecast combination procedure called AI-AFTER that will perform well given the data and candidate forecasts at hand. Detailed simulation studies using variants of the data scenarios shown in Examples 1 and 2 as well as time series examples with our AI-AFTER proposal will be given in Section 5; we will also visit the two simulations of Section 2.2 again in Section 5.4 to verify that the proposed AI-AFTER may perform well simultaneously whether the underlying data scenario favors the objective of adaptation or improvement.

3. Testing the Potential of Combining for Improvement

3.1. Our approach

We approach the potential of improvement from a hypothesis testing framework. As discussed before, the choice to combine for adaptation is a relatively conservative strategy. Therefore, we place CFA in the role of the null hypothesis and CFI as the alternative. The test recommends combining for improvement if the data provide evidence that CFI is a potentially useful strategy.

The choice to combine for improvement makes sense if, for the available n, $R(\Psi_L; n)$ is significantly less than $R(\Psi_0; n)$ so that some linear combination of the original forecasters could outperform all of the original forecasters. The test we propose in this section estimates these two risks and measures the evidence for improvement potential. For a given class Ψ, $R(\Psi; n) = R(\psi^*; n)$; therefore, estimation of $R(\Psi; n)$ involves identifying the best procedure ψ^* and estimating its risk. Our estimated comparison of $R(\Psi_0; n)$ and $R(\Psi_L; n)$ then involves comparing the best-performing original forecaster to the best-performing CFI combination procedure.

For the class Ψ_0, it is possible to estimate the risk of each original forecaster under squared error loss by calculating its mean squared forecast error (MSFE) on the sample data at hand. The minimum of these estimated risks then serves as an estimate of $R(\Psi_0; n)$. Practical estimation of $R(\Psi_L; n)$ requires a consideration of what forms of linear combinations one wants to consider. Although previous authors have derived asymptotically optimal weights in Ψ_L under certain conditions, for finite samples the optimal weights are typically unknown and many combinations in Ψ_L might be reasonable to try. Typical examples include least squares or penalized linear regression of the response on the original procedures, bagging (Breiman, 1996), boosting (Freund and Schapire, 1995; Yang et al., 2018), and/or even taking a simple average of the forecasts.

Clearly, it is impossible to analyze all of the linear combinations in Ψ_L. To estimate the risk $R(\Psi_L; n)$, we can only evaluate a finite subset $\Psi \subset \Psi_L$. We want Ψ to be large enough to allow exploration of different ways to combine the forecasts for improvement. We specifically suggest two ways of approaching the size of Ψ in the AI-AFTER testing procedure. First, one can make $|\Psi|$ approximately equal to M, the original number of candidate forecasts. By making the two sets of forecasters (original and combined) similar in size, we

mitigate the chance of the multiple comparison phenomenon giving an unfair advantage to either set when we compare the empirically best performer in each. One way to generate about M combined forecasts from the M original forecasts (when $n > M$) is to consider linear regressions of the response on an intercept and a sequence of $\{1, \ldots, M\}$ of the original forecasts chosen via a forward or backward selection algorithm (e.g., Zhang, 2011; Ing and Lai, 2011; Qian et al., 2019a). Second, one may wish to include many more than M combined forecasts in Ψ in order to search for one combination procedure that works well. In this case, we describe a randomization test in Step 6′ of Section 3.2 to correct for the increased chance of a Type I error due to multiple comparisons, as well as a safeguard feature to be described in Section 4.3.

3.2. A testing procedure

We next describe a procedure to determine the potential of improvement given the available data and the candidate forecasts. We assume the availability of observed responses $Y_1, \ldots, Y_n$ and a corresponding $n \times M$ matrix X containing the forecasts of the n responses from the M candidates.

1. Select a fraction ρ of observations, and the initial $n_0 = \lfloor \rho n \rfloor$ observations will only be used to estimate combining weights and build the combined forecasts for later observations (to be defined in Step 2).
2. Decide on a set Ψ of combined forecasting procedures from one or more families of forecast combination procedures that combine for improvement. Examples of combination procedures to consider putting in Ψ will be given later in this section. Let $\widetilde{M} = |\Psi|$ denote the number of combination methods in Ψ, and denote the particular combination procedures in Ψ as $\psi_1, \ldots, \psi_{\widetilde{M}}$.
3. For every observation $i = n_0 + 1, \ldots, n$, calculate $\widetilde{M}$ forecasts of Y_i by

applying the combination methods in Ψ to the first $i-1$ observations (responses and original candidate forecasts).

4. The previous step produces $\widetilde{M}$ combined forecasts for each of the most recent $n_1 = (n - n_0)$ observations. For each $j \in \{1, \ldots, \widetilde{M}\}$, compute the empirical (pseudo-out-of-sample) MSFE for ψ_j over the most recent n_1 observations. Identify the procedure $\psi^* \in \Psi$ with the minimum empirical MSFE.
5. Similar to the previous step, identify the original candidate procedure $\delta^* \in \Delta$ with the minimum empirical MSFE over the most recent n_1 observations.
6. Compare the forecast errors of ψ^* to the forecast errors of δ^* using the D-M test (Diebold and Mariano, 1995) under squared error loss. The null hypothesis of the test is that the procedures ψ^* and δ^* are equally accurate in forecasting. Testing the potential of improvement suggests a one-sided alternative hypothesis that ψ^* is more accurate than δ^*.

6′. A detailed study of the null distribution of the D-M test statistic in this framework is beyond the scope of this paper. However, when the number of combined forecasting procedures is much larger than the number of original candidate forecasts (that is, when $\widetilde{M} \gg M$), the null distribution of the p-value from Step 6 may be shifted toward zero due to the multiple comparison phenomenon. We recommend correcting for this by performing a randomization test that uses simulation to approximate the null distribution of the p-value from Step 6. The randomization test works as follows:

   (a) Create an $n_1 \times (M + \widetilde{M})$ matrix F of the original and combined forecasts for the most recent n_1 observations.
   (b) Do each of the following steps N times, where N is a large number

of repetitions for randomization:

i. Randomly permute the $M + \widetilde{M}$ columns of F. Label the first M columns of the permuted matrix as the "original" forecasts and the remaining $\widetilde{M}$ columns as "combined" forecasts.

ii. Perform Steps 4-6 on the sets of "original" and "combined" forecasts.

This step produces a set of N randomization p-values. These are D-M p-values under random labeling of the forecasting procedures as "original" and "combined".

(c) The p-value of the randomization test is the proportion of randomization p-values from Step 6′(c) that are less than or equal to the p-value observed in Step 6.

7. Compare the p-value from the D-M test to a pre-specified significance level α. A small p-value indicates evidence that at least one of the combined forecasting procedures in Ψ is more accurate than the best original forecaster in Δ, giving empirical evidence that $R(\Psi_L; n) < R(\Psi_0; n)$ and indicating the potential of improvement using the available data.

Step 2 of the AI-AFTER testing procedure is agnostic regarding the nature and the number of the combination procedures that can be included in Ψ. In practice, we construct Ψ from the following combinations of the M original forecasts.

- **LASSO with AIC, BIC.** The LASSO solutions with the minimum values of AIC and BIC on the solution path are found using the R package `ncvreg`.

- **Stepwise selection with AIC, BIC.** A stepwise selection algorithm is applied to the first i observations, and the least squares models with the minimum AIC and BIC on the solution path are chosen.

- **Best subset of each size.** The initial n_0 observations are used to select the best model of each size from 1 to $\min(M, n_0 - 1)$. If $\min(M, n_0 - 1) \leq 20$, an exhaustive search is done to find the best model of each size; otherwise, this is done via forward selection. The coefficients of these models are updated after every observation i, but the set of active variables in each model stays the same.

- **Constrained linear regression (CLR).** We consider a constrained linear regression (CLR) of the original forecasters without an intercept. Specifically,

$$\hat{y}_i^{\text{CLR}} = \sum_{j=1}^{M} \hat{\beta}_{i,j} \hat{y}_{\delta_j,i}, \text{ where}$$

$$\left\{\hat{\beta}_{i,1}, \ldots, \hat{\beta}_{i,M}\right\} = \underset{\beta_1,\ldots,\beta_M}{\operatorname{argmin}} \sum_{t=1}^{i-1} \left(Y_t - \sum_{j=1}^{M} \beta_j \hat{y}_{\delta_j,t} \right)^2,$$

  with the constraints that $\hat{\beta}_{i,j} \geq 0$ for $1 \leq j \leq M$ and $\sum_{j=1}^{M} \hat{\beta}_{i,j} = 1$.

- **Bates-Granger ($\mathbf{BG_{0.9}}$, $\mathbf{BG_1}$).** Under the weighting scheme described in Bates and Granger (1969), we have

$$\hat{y}_i^{\text{BG}} = \sum_{j=1}^{M} w_{i,j} \hat{y}_{\delta_j,i}.$$

  The combining weights are

$$w_{i,j} = \frac{\left(\hat{\sigma}_{i,j}^2\right)^{-1}}{\sum_{k=1}^{M} \left(\hat{\sigma}_{i,k}^2\right)^{-1}}, \text{ where}$$

$$\hat{\sigma}_{i,k}^2 = \frac{1}{i-1} \sum_{t=1}^{i-1} \rho^{i-(t+1)} \left(Y_t - \hat{y}_{\delta_i,t}\right)^2 \text{ with } \rho = 0.9 \text{ or } 1.$$

- **Naïve combinations (SA, MD, TM).** We consider three naïve forecast combinations that do not take individual forecaster performance into account:

– SA, the simple average of the candidate forecasts $\hat{y}_{\delta_j,i}, 1 \leq j \leq M$.

– MD, the median of the candidate forecasts.

– TM, the trimmed mean. At each step i, the highest and lowest $\lfloor M/20 \rfloor$ values of $\hat{y}_{\delta_j,i}$ are removed before taking the mean.

In the above construction of Ψ, we have up to $\widetilde{M} = \min(M, n_0 - 1) + 10$ distinct combinations of the original forecasts at each step i (It is possible for some of the combination methods to choose the same forecasting procedures; for example, the subsets chosen by AIC and BIC could match). The AI-AFTER testing procedure described in this section provides helpful information about the nature of the forecasting problem for the data at hand, and indicates the potential of improvement to enable the choice of a proper forecast combination method to be discussed next.

4. Combining for Adaptation or Improvement

4.1. Using AFTER algorithm

The hypothesis test described in Section 3 indicates whether given the data and candidate forecasts at hand, one should perform combining for improvement. If there is little evidence that any combined forecast can outperform the best individual forecast, then one may simply target the forecast risk of the best individual forecaster. As AFTER can provides protection against model selection uncertainty (Zou and Yang, 2004) and automatically adapt to changes over time in the data generating process and relative forecaster performance, we adopt AFTER in scenarios favorable to the objective of adaptation; see also Section 2.1 for a brief description of the AFTER method.

If the hypothesis test indicates a combining for improvement scenario, this means we were able to generate one or more forecast combinations that outper-

form all of the original forecasters. In this case, our goal should be to match the performance of the best combined forecast. In other words, this can be thought of as an adaptation scenario in which the combined forecasters are considered as the candidates. Framing the problem in this way suggests that we need to propose a new version of AFTER to achieve the potential for improvement by using the combined forecasters instead of the original ones as the forecast candidates for AFTER.

4.2. AI-AFTER algorithm

The preceding discussion suggests that AFTER, applied to the proper forecasting procedures, can be used to effectively combine forecasts for adaptation (A) or improvement (I), and we call this new version AI-AFTER. In the following, we summarize AI-AFTER, where the hypothesis test described in Section 3 will be informative on the direction. Recall that we assume the analyst starts with observed responses $Y_1, \ldots, Y_n$ and a corresponding $n \times M$ matrix of forecasts X.

1.-7. Apply the AI-AFTER testing procedures described in Section 3. The p-value from the hypothesis test is used to determine which set of forecasts is combined by AFTER in the final step.

8. If the p-value is greater than the level α, we conclude that CFA is an appropriate strategy and therefore apply the AFTER algorithm to the original responses Y and forecasts X. Otherwise, we apply AFTER to the $n_1 \times \widetilde{M}$ matrix $\widetilde{X}$ of combined forecasts (and the corresponding most recent n_1 observed responses) produced in Step 3 of the testing procedures. The forecasts $\hat{y}_{1,i}$'s ($i = n_0 + 1, \cdots, n$) can then be obtained according to the determined combining direction, and the corresponding combining weight vector (in $\mathbb{R}^M$ or $\mathbb{R}^{\widetilde{M}}$) is generated for next forecasting outcome.

For practical use of AI-AFTER, the performance of AFTER often benefits from using the first few n_{init} records as "burn-in" observations. This means the first n_{init} forecasts produced by AFTER assign equal weights to all candidates, since relative forecaster performance cannot be reliably estimated using only the first few observations. The n_{init} initial observations are used only to estimate the combining weights for future observations, and we set $n_{\text{init}} = 5$ by default.

4.3. A safeguard feature

In the previous subsection, the hypothesis test is integrated as an informative step to decide on the direction of combining (CFA or CFI) to potentially lead to improved forecasting performance. On the other hand, since hypothesis test is subject to type I and/or type II errors (particularly when sample size is small with relatively low power), in the following, we further introduce a safeguard feature into AI-AFTER to protect against these possible errors, without necessarily making compromise on forecasting performance (to be discussed in Section 4.4). Specifically, we devise a composite AFTER step: AFTER is first applied to an expanded set of original and combined forecasters to generate new "safeguard" forecasts $\hat{y}_{2,i}$'s; subsequently, an extra layer of AFTER is applied by treating $(\hat{y}_{1,i}, \hat{y}_{2,i})$ as two candidate forecasters and generate the corresponding combining weights for forecasts $\hat{y}_{C,i}$'s. This safeguard step is summarized as follows.

9. Create an $n_1 \times (M + \widetilde{M})$ matrix $\widetilde{\widetilde{X}}$ of candidate forecasts by concatenating $\widetilde{X}$ to the most recent n_1 rows of X. Apply the AFTER algorithm to the most recent n_1 responses and the matrix $\widetilde{\widetilde{X}}$ consisting of the M original forecasters and the $\widetilde{M}$ combined forecasters; this generates the "safeguard" forecasts $\hat{y}_{2,i}$'s $(i = n_0 + 1, \cdots, n)$ and combining weight vector in $\mathbb{R}^{M+\widetilde{M}}$ for next forecasting outcome.

10. Apply the AFTER algorithm again with $(\hat{y}_{1,i}, \hat{y}_{2,i})$ as candidate forecasts. This creates forecasts $\hat{y}_{C,i}$'s and a combining weight vector in $\mathbb{R}^2$ to use for forecasting the next outcome.

4.4. Risk bound of AI-AFTER

Recall that we have $\delta_1, \cdots, \delta_M$ as the original candidate forecasting procedures, producing $\hat{y}_{\delta_j,i}$, where $i = 1, 2, \cdots$, and $1 \le j \le M$. In the construction of Ψ, we have $\widetilde{M} = |\Psi|$ distinct combined forecasts, and let κ_l $(1 \le l \le \widetilde{M})$ denote these forecast combination procedures. To show forecast risk bound of the proposed AI-AFTER approach, we assign prior probabilities: fix some constant p_0 $(0 < p_0 < 1)$ so that each δ_i $(i = 1, \cdots, M)$ has prior $\pi = (1 - p_0)\frac{1}{M}$ and each κ_l $(l = 1, \cdots, \widetilde{M})$ has prior $\pi = \frac{p_0}{\widetilde{M}}$. Note that the above prior probabilities add up to 1. Let N be a total forecasting horizon. Suppose Conditions 6 and 8 in Yang (2004) are satisfied. Then we immediately have the final combined forecasts of AI-AFTER as shown in Theorem 1 that satisfy near optimal performance.

Theorem 1. Let δ_C be the AI-AFTER combining procedure. Then there is some constant λ from (2) such that $\sum_{i=n+1}^{N} E(Y_i - \hat{y}_{\delta_C,i})^2$ is no larger than the smallest of the following:

$$\begin{cases} \log(M) + \log \frac{2}{1-p_0} + \inf_{1 \le j \le M} \sum_{i=n+1}^{N} E(Y_i - \hat{y}_{\delta_j,i})^2, \\ \log(\widetilde{M}) + \log \frac{2}{p_0} + \inf_{1 \le l \le \widetilde{M}} \sum_{i=n+1}^{N} E(Y_i - \hat{y}_{\kappa_l,i})^2. \end{cases}$$

From Theorem 1, without any prior knowledge and in a universal fashion, AI-AFTER is no worse than the (unknown) best original individual forecasts in Ψ_0 and the best of the combined forecasts in Ψ, plus relatively small additive penalty (when n is large). Consequently, if the average squared forecast error does not converge to 0, we have that $\overline{\lim}_{N\to\infty} \frac{\sum_{i=n+1}^{N} E(Y_i - \hat{y}_{\delta_C,i})^2}{R} \le 1$, where R

is the minimum of $\inf_{1\le j\le M}\sum_{i=n+1}^{N}E(Y_i-\hat{y}_{\delta_j,i})^2$ and $\inf_{1\le l\le \widetilde{M}}\sum_{i=n+1}^{N}E(Y_i-\hat{y}_{\kappa_l,i})^2$. Therefore, AI-AFTER is asymptotically no worse than the best of original forecasts and the combined forecasts, and is adaptively intelligent: it is conservative (achieving combining for adaptation) when there is no advantage in pursuing improvement, and it is aggressive otherwise, achieving the good performance of alternative combined forecasts (such as sparse regression combining) for improvement.

5. Simulation studies

In this section we present simulation results for a linear regression setting and for a time series setting, followed by re-visiting the two simulations of Sections 2.2. In the linear regression setting, a large number of covariates help to determine the data generating process and are considered by M different candidate models. In the time series setting, past values of the response variable are used in the candidate models in addition to one or two covariates. In both settings, we present forecast combination scenarios of adaptation or improvement. Throughout the simulation, ρ is set to $1/3$, and a level of $\alpha = 0.1$ is used. We compare the predictive performance of AI-AFTER to five competitors: the basic AFTER method; the methods of CLR, BG_1 and SA described in Section 3.2; and combination via linear regression (LR).

In each simulation setting described below, 200 independent realizations of data are generated. Each realization includes n_{train} rows of training data (that are used only to build the candidate forecasts and are not available to the analyst), the n observations and candidate forecasts (available to the analyst tasked with combining the forecasts), and a subsequent number n_{eval} of outcomes used to evaluate the combining methods. In all of our simulations, we set $n_{\text{train}} = n$ and $n_{\text{eval}} = 20$. For each realization, the combining methods are evaluated by

their mean squared forecast error (MSFE) over these n_{eval} observations. Specifically, the MSFE for a combining method Δ in a given realization j is

$$\text{MSFE}_j^{\Delta} = \frac{1}{20} \sum_{i=n+1}^{n+20} \left(y_{i,j} - \hat{y}_{i,j}^{\Delta} \right)^2, \tag{9}$$

where $y_{i,j}$ denotes the value of the ith observation in the jth realization, and $\hat{y}_{i,j}^{\Delta}$ is the forecast of $y_{i,j}$ produced by the application of method Δ. In each setting, we show summaries of $\text{MSFE}_j^{\Delta}/\text{MSFE}_j^{\text{SA}}$ for the different methods Δ to compare the performance of each combination method relative to a simple average of the candidate forecasts.

5.1. Linear regression examples

In this section, the true model is

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{10}$$

where the $\mathbf{x}_{i\cdot} = (x_{i1}, \cdots, x_{ip})$ are i.i.d. multivariate normal with mean $\mathbf{0}$ and covariance matrix elements $\sigma_{jk} = 0.5^{|j-k|}$, $p = 30$, $\beta = (3, 2, 1, 1, 1, 1, 0, \ldots, 0)$, and the ε_i are i.i.d. $N(0, 4)$ and independent of $\mathbf{x}_{i\cdot}$. We consider both an adaptation scenario and an improvement scenario (denoted by OLS-Adaptation and OLS-Improvement, respectively); the data generating process is the same (10) in both settings, but the candidate forecasts available to the analyst differ.

5.1.1. OLS-Adaptation scenario

The candidate forecasts in the adaptation scenario, for $1 \leq i \leq n$, are constructed as follows:

$$\begin{aligned} \text{Forecast } 1 : \hat{y}_{i,1} &= \hat{\beta}_{1,0} + \sum_{j=1}^{6} \hat{\beta}_{1,j} x_{ij}, \\ \text{Forecast } k,\ 2 \leq k \leq 30 : \hat{y}_{i,k} &= \hat{\beta}_{k,0} + \sum_{j=1}^{p} I(k,j) \hat{\beta}_{k,j} x_{ij}. \end{aligned} \tag{11}$$

The $\hat{\beta}_{k,j}$ are the least squares estimates, trained using the first $i-1$ observations available to the analyst as well as a previous set of $n_{\text{train}} = n$ historical observations of $(y, \mathbf{x})$ that are used to train the candidate forecasts only and are not available to the analyst. The $I(k, j)$ are independent Bernoulli(0.5) random variables; therefore, each covariate has a 50/50 chance of being included as a predictor variable in each candidate forecast k, $2 \leq k \leq 30$. Since none of the forecasters k for $k > 1$ add any information about the data generating process to (11), in this case $R(\Psi_0; n) = R(\Psi_L; n)$ for large enough n. Thus, the above set of candidate forecasters represents a scenario where combining for adaptation is the appropriate goal.

5.1.2. OLS-Improvement scenario

Now for the same data generating process (10), consider candidate forecasts $k, 1 \leq k \leq 30$, of the form

$$\text{Forecast } k : \hat{y}_{i,k} = \hat{\beta}_{k,0} + \hat{\beta}_k x_{ik}. \tag{12}$$

Again the $\hat{\beta}_k$ are trained using the previous $n + i - 1$ observations. In this scenario, each forecaster on its own has incomplete information about the data-generating process, but if the forecasts are combined in a smart way, the combination can capture all of the information in (10) and thus produce a more accurate forecast.

5.2. Time series examples

We consider an autoregressive (AR) process with up to two covariates. In contrast to the linear regression examples considered previously, both the adaptation and improvement cases consider the same set of candidate forecasts. The nature of the data generating process determines whether combining for adaptation or improvement is more appropriate. The candidate forecasters are assumed to have access to the previous response values and at most one of the two

covariates. Specifically, each candidate forecast is based on an AR(X) model according to the following:

$$\text{Forecast 1: } \hat{y}_{i,1} = \hat{\beta}_{0,1} + \hat{\gamma}_{1,1} y_{i-1}$$

$$\text{Forecast 2: } \hat{y}_{i,2} = \hat{\beta}_{0,2} + \hat{\gamma}_{1,2} y_{i-1} + \hat{\gamma}_{2,2} y_{i-2}$$

$$\text{Forecast 3: } \hat{y}_{i,3} = \hat{\beta}_{0,3} + \hat{\beta}_{1,3} x_{i1} + \hat{\gamma}_{1,3} y_{i-1}$$

$$\text{Forecast 4: } \hat{y}_{i,4} = \hat{\beta}_{0,4} + \hat{\beta}_{2,4} x_{i2} + \hat{\gamma}_{1,4} y_{i-1}$$

$$\text{Forecast 5: } \hat{y}_{i,5} = \hat{\beta}_{0,5} + \hat{\beta}_{1,5} x_{i1} + \hat{\gamma}_{1,5} y_{i-1} + \hat{\gamma}_{2,5} y_{i-2}$$

$$\text{Forecast 6: } \hat{y}_{i,6} = \hat{\beta}_{0,6} + \hat{\beta}_{2,6} x_{i2} + \hat{\gamma}_{1,6} y_{i-1} + \hat{\gamma}_{2,6} y_{i-2}$$

As in Section 5.1, the $\hat{\beta}$ and $\hat{\gamma}$ coefficients are trained on the first $n + i - 1$ observations of $(y, \mathbf{x})$, with the first n observations of the series unavailable to the analyst. The values of $\mathbf{x}_{i\cdot}$ are i.i.d. with two independent standard normal covariates. We next present both an adaptation scenario and an improvement scenario (denoted by AR(X)-Adaptation and AR(X)-Improvement, respectively).

5.2.1. AR(X)-Adaptation

We apply the six candidate forecasts described above to predict a y_i generated by the following process:

$$y_i = 0.5 y_{i-1} + 0.4 y_{i-2} + \varepsilon_i, \tag{13}$$

where the ε_i are i.i.d. $N(0, 4)$ as in Section 5.1. In this scenario, Forecast 2 represents the true model and has the lowest forecasting risk (for large enough n). Forecasts 5 and 6 include both AR lags, but each also uses one non-informative covariate. Forecasts 1, 3 and 4 fail to include the second lag of y in their model. For large enough n, the performance of Forecast 2 cannot be improved by combining it with the other forecasters; therefore, combining for adaptation is considered more appropriate.

5.2.2. AR(X)-Improvement

The data-generating process uses both covariates:

$$y_i = 0.5y_{i-1} + 0.4y_{i-2} + x_{i1} + x_{i2} + \varepsilon_i. \tag{14}$$

The ε_i are again i.i.d. $N(0, 4)$. While none of the six candidate forecasts use both covariates, some use x_1 and some use x_2. Therefore, forecast combination can result in improved performance over any individual forecaster due to sharing of information. Thus, combining for improvement is considered more appropriate.

5.3. Results

5.3.1. AI-AFTER test

We first evaluate the performance of the AI-AFTER testing procedure described in Section 3 for determining the combining goal.

Table 1: Percentage of 200 realizations that AI-AFTER selected Combining for Improvement as the proper goal of forecast combination.

Data-Generating Process (DGP)	Sample Size	% Rejected H_0
OLS-Adaptation	$n = 100$	1.0%
	$n = 300$	2.5%
AR(X)-Adaptation	$n = 100$	0.5%
	$n = 300$	1.0%
OLS-Improvement	$n = 100$	100.0%
	$n = 300$	100.0%
AR(X)-Improvement	$n = 100$	39.5%
	$n = 300$	99.0%

Table 1 shows, for each of the four simulation settings at sample size levels $n = 100$ and $n = 300$, the proportion of the 200 realizations that the test rejected H_0 and recommended combining for improvement.

Figure 3: Linear Regression examples: p-value distribution for AI-AFTER test.

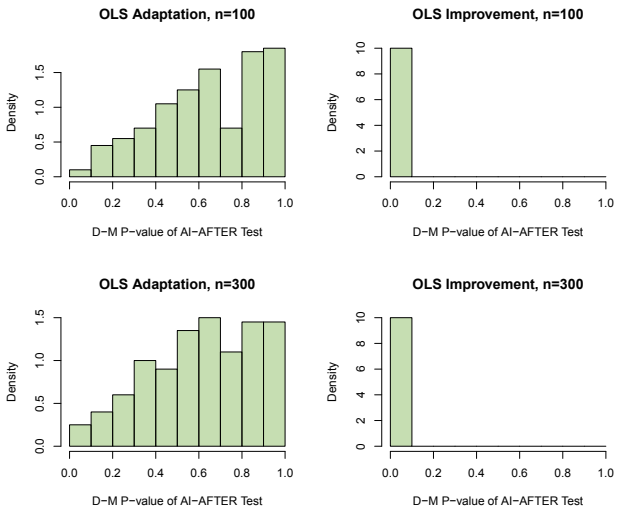
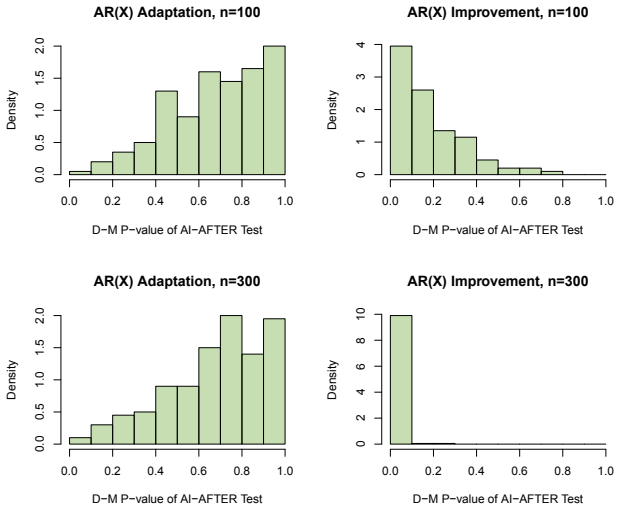


Figure 4: AR(X) examples: p-value distribution for AI-AFTER test.



The top half of the table shows that in all cases, the frequency of Type I errors was lower than the nominal $\alpha = 0.1$. This could be understood by the observation that $H_0 : R(\Psi_0; n) = R(\Psi_L; n)$ is not true; instead, in each

case $R(\Psi_0; n) < R(\Psi_L; n)$ because the expected performance of the true model, which is among the candidates, is better than the expected performance of any combination method in Ψ_L. In each of the four scenarios where combining for adaptation is appropriate (and combining for improvement carries additional risk with no reward), the test rejected H_0 and recommended combining for improvement less than 3% of the time.

The bottom half of Table 1 shows that the AI-AFTER test does well in discovering the potential of improvement when $n = 300$, with H_0 rejection rates of 100% and 99.0% in the OLS- and AR(X)-Improvement scenarios, respectively. As with most hypothesis tests, the test is less powerful when n is smaller. For example, when $n = 100$ in the combining for improvement cases, the results were mixed, with H_0 rejected 100.0% of the time in the OLS setting but 39.5% of the time in the AR(X) setting. These results can also be observed from the p-value distributions as shown in Figures 3 and 4.

The results in these examples suggest that the test is effective at controlling Type I error when the true data generating process is represented in the candidates and thus combining for adaptation is appropriate. When combining for improvement is the proper goal, the test was most effective and informative at larger sample sizes such as $n = 300$, but may exhibit lower power (as expected for any tests) if we use smaller sample size. We next examine the robustness of AI-AFTER forecasting performance with the integrated test.

5.3.2. Forecasting Performance

Table 2 compares the forecasting performance of AI-AFTER against AFTER, BG_1, LR, CLR, and SA. For each method Δ, the table shows averages and standard errors of the ratio of $MSFE_j^{\Delta}$ to $MSFE_j^{SA}$. Figures 5 and 6 show the empirical distributions of the MSFE (relative to SA) for each method over

Table 2: Performance comparison: Mean (S.E.) of $\text{MSFE}_j^{\Delta}/\text{MSFE}_j^{\text{SA}}$ for each combination method Δ over 200 realizations.

			AI-AFTER	AFTER	BG_1	LR	CLR
OLS DGP	CFA	$n = 100$	0.795 (0.008)	0.792 (0.008)	0.883 (0.004)	0.909 (0.011)	0.802 (0.007)
		$n = 300$	0.779 (0.007)	0.777 (0.007)	0.875 (0.003)	0.806 (0.008)	0.778 (0.007)
	CFI	$n = 100$	0.386 (0.005)	0.735 (0.008)	0.960 (0.001)	0.419 (0.006)	0.637 (0.005)
		$n = 300$	0.365 (0.005)	0.749 (0.008)	0.960 (0.001)	0.379 (0.006)	0.633 (0.005)
AR(X) DGP	CFA	$n = 100$	0.981 (0.003)	0.978 (0.003)	0.996 (0.001)	1.013 (0.006)	0.978 (0.003)
		$n = 300$	0.983 (0.003)	0.983 (0.003)	0.997 (0.001)	0.995 (0.004)	0.983 (0.003)
	CFI	$n = 100$	0.915 (0.007)	0.984 (0.007)	0.990 (0.001)	0.908 (0.009)	0.940 (0.004)
		$n = 300$	0.878 (0.007)	0.998 (0.007)	0.989 (0.001)	0.880 (0.007)	0.937 (0.004)

Figure 5: Linear Regression examples: Each boxplot shows $\text{MSFE}_j^{\Delta}/\text{MSFE}_j^{\text{SA}}$ for 200 realizations.

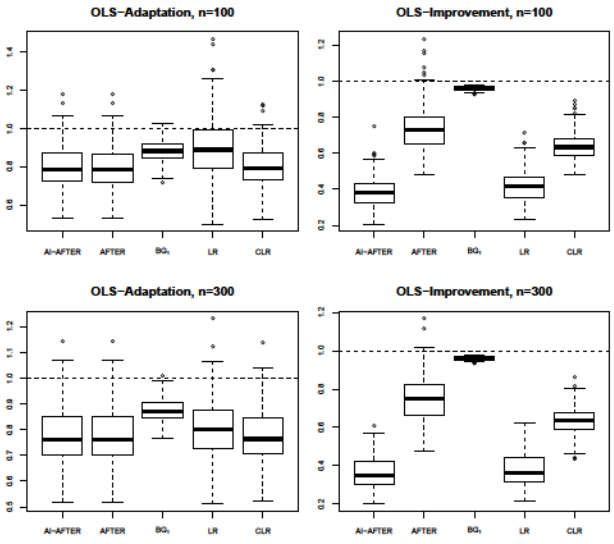
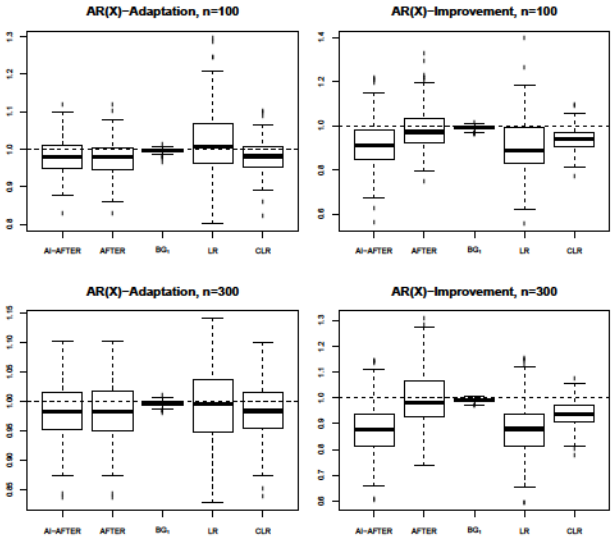


Figure 6: AR(X) examples: Each boxplot shows $\text{MSFE}_j^{\Delta}/\text{MSFE}_j^{\text{SA}}$ for 200 realizations.



the 200 realizations.

In alignment with our expectation, AI-AFTER is able to closely track AFTER in the adaptation settings. In the improvement settings, AI-AFTER remains to perform competitively against other forecast combination methods at different sample sizes considered; it offers significant improvement over AFTER because AFTER is not designed for the goal of improvement. Overall, the simulation results show that AI-AFTER features a forecast combination strategy that is aggressive when the reward from combining for improvement is high and conservative when a combination of forecasts cannot improve much (if at all) over a single outstanding candidate.

5.4. Simulations 1 and 2 re-visited

We next re-visit the two simulations of Section 2.2 to verify that the dilemma observed from the simulation experiment in Figures 1 and 2 can now be solved. The AI-AFTER algorithm is applied to the same simulation experiment of Section 2.2, and the averaged relative MSFEs of Simulation 1 and Simulation 2 are summarized in Figure 7 and Figure 8, respectively. Satisfactorily, Figure 7 shows that in Simulation 1, AI-AFTER performs very similarly to AFTER that pursues the CFA objective; in contrast, Figure 8 shows that in Simulation 2, AI-AFTER performs almost as well as LR when the signal is relatively large, and maintains performance similar to BG and SA without incurring the excessive cost of LR when the signal becomes weak. AI-AFTER indeed performs well simultaneously under both data scenarios with different favorable combining objectives.

Figure 7: Relative forecast performance of AI-AFTER in Example 1.

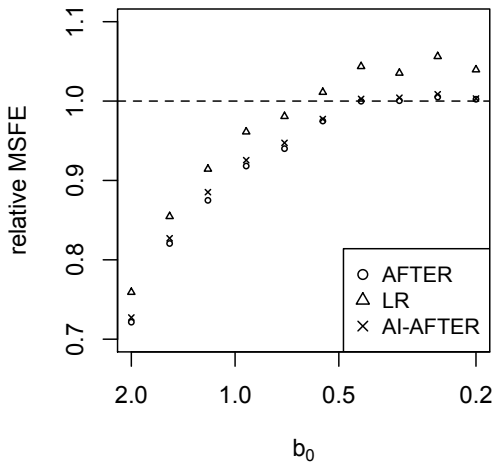
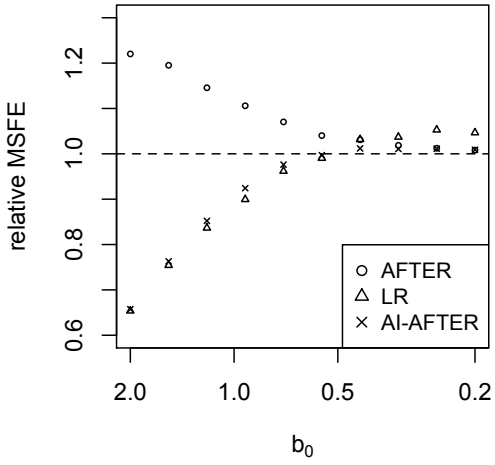


Figure 8: Relative forecast performance of AI-AFTER in Example 2.



6. Output Forecasting

6.1. Data, Forecasts, and Combining Methods

We next apply the method of AI-AFTER to forecast two measures of output growth for seven developed countries using data first analyzed in Stock and Watson (2003). Specifically, we forecast

$$Y_{t+4h} = \frac{100}{h} \ln(Q_{t+4h}/Q_t),$$

where, depending on the analysis, Q is either a country's real GDP (RGDP) or Index of Industrial Production (IP), t represents the current quarter at the time of forecasting, and h represents the forecasting horizon in terms of number of years ahead. We consider forecasts for $h = 1$ and 2-year horizons. For each horizon, there are 13 forecasting problem cases considered: RGDP and IP for each of the seven countries, except IP for France (data not available for enough periods).

Following Stock and Watson (2003), the data was used to study the effectiveness of individual asset prices as leading indicators of output growth. Consider

forecasting models of the form

$$Y_{t+4h} = \beta_0 + \beta_1(\mathrm{L})X_t + \beta_2(\mathrm{L})Y_t + u_{t+4h}, \tag{15}$$

where u_{t+4h} is an error term and $\beta_1(\mathrm{L})$ and $\beta_2(\mathrm{L})$ are lag polynomials allowing multiple lagged values of X and Y to be included in the regression; the X variables include interest rates, exchange rates, stock and commodity prices, and other measures; a full list of X variables considered for each country can be found in Stock and Watson (2003). For each country, up to 73 different candidate predictors X_t are used, one at a time, in the model form (15) to predict Y_{t+4h}. The data series are recorded quarterly for each country from 1959 to 1999. For each forecasting problem, the first 50 available observations are used to train the candidate forecasts in (15) and are considered unavailable to the combining analyst. After these restrictions, the number of valid h-step-ahead responses ranges from 82 to 100, while the number of individual candidate forecasts of the form (15) ranged from 26 to 64, depending on the availability of data series for each country. Forecast combination methods are employed to generate combined forecasts.

As in Section 5, we compare the forecasting performances of AI-AFTER, AFTER, BG_1, LR, CLR, and SA. The accuracy of each combination method Δ, in terms of $\mathrm{MSFE}^{\Delta}/\mathrm{MSFE}^{\mathrm{SA}}$, over the final $n_{\mathrm{eval}} = 20$ values of response outcomes is recorded. We set $\alpha = 0.1$ and treat the final 20 outcomes as being unavailable to the analyst, so they are not used in the calculation of the AI-AFTER testing procedure's p-value to determine the direction of combining.

6.2. Results

The relative performance of each forecast combination method for predicting growth in RGDP and IP can be found in Table 3 and Table 4, respectively. There are 13 sets of forecasts (six countries for RGDP, seven countries for IP)

Table 3: MSFEs of combination forecasts, relative to SA: Forecasts of h-year growth of RGDP.

	Canada	Germany	Italy	Japan	UK	USA
$h = 1$						
AI-AFTER p-value	0.007	0.761	0.005	10^{-5}	0.201	0.079
AI-AFTER	1.067	0.535	0.663	0.776	0.904	1.443
AFTER	1.439	2.705	0.756	1.663	1.530	6.077
BG_1	1.006	1.024	0.952	0.941	0.925	1.001
LR	1.119	0.714	0.645	0.757	1.422	2.424
CLR	1.095	0.781	0.651	1.206	1.082	1.433
$h = 2$						
AI-AFTER p-value	0.095	0.005	10^{-6}	10^{-5}	10^{-5}	0.064
AI-AFTER	0.628	0.407	0.177	0.456	1.081	0.790
AFTER	1.826	1.595	0.438	0.955	1.412	7.078
BG_1	1.077	0.784	0.781	0.815	0.721	1.048
LR	1.601	0.404	0.265	0.465	1.006	1.379
CLR	1.097	0.199	0.185	0.866	0.772	1.911

at two horizons, for a total of 26 cases. For each forecast horizon, the tables show the p-values of the AI-AFTER tests, as well as the MSFEs of each forecast combination method, relative to the combination by SA.

First, except for only five cases, the forecasting performance of original AFTER method is not satisfactory even compared to LR, which seems to suggest that our considered forecasting scenarios from these data sets should overall belong to the forecast improvement category. In particular, setting level $\alpha = 0.1$, the AI-AFTER tests identify 21 cases to be combining for improvement. For the remaining 5 cases, their insignificance results could be attributed to the type II error (as observed from our simulation studies) under the scenarios with rela-

Table 4: MSFEs of combination forecasts, relative to SA: Forecasts of h-year growth of IP.

	Canada	France	Germany	Italy	Japan	UK	USA
$h = 1$							
AI-AFTER p-value	0.013	0.035	0.097	0.006	0.005	0.048	0.100
AI-AFTER	0.630	0.736	0.875	0.906	0.729	0.738	0.644
AFTER	2.227	0.967	2.055	1.198	1.069	0.931	1.459
BG_1	0.958	0.990	1.053	0.976	0.943	0.992	0.941
LR	2.463	1.081	0.259	0.873	0.659	0.808	0.882
CLR	0.646	0.749	1.513	0.856	0.799	0.542	0.627
$h = 2$							
AI-AFTER p-value	0.452	0.473	0.093	0.007	0.004	0.001	0.326
AI-AFTER	0.932	0.508	0.953	0.542	0.409	0.815	0.579
AFTER	4.086	0.514	1.278	0.966	1.517	0.337	2.314
BG_1	0.862	0.720	0.738	0.780	0.745	0.947	0.794
LR	3.204	0.567	1.822	0.965	0.323	0.649	0.671
CLR	0.997	0.283	0.825	0.540	0.697	0.968	0.272

tively small sample size; nevertheless, AI-AFTER remains to give better (or at least as good as) MSFE results than that of AFTER, confirming that the safeguard feature of AI-AFTER indeed takes effect to ensure desirable forecasting performance.

These observations above for forecast improvement may not be surprising: Stock and Watson (2003) found that none of the individual forecasts of the form (15) performed reliably well over the entire analysis period; after comparing these individual forecasts to those from a benchmark AR model in two separate time periods, they found that "forecasting models that outperform the AR in the first period may or may not outperform the AR in the second,

but whether they do appears to be random". Due to the lack of any reliably outstanding individual predictor(s), their results imply that for these data and forecast candidates, the best individual forecaster can be improved by forecast combination; our results above are largely consistent with their findings, which find no consistently accurate individual forecasts of the form (15) but do find success by combining these forecasts (for improvement). On the other hand, Stock and Watson (2004b) also found that simple average of the individual forecasts was often reliably accurate for these forecasting problems, and supported the strategy of simple forecast combination with little or no time variation in the combining weights. However, our analysis finds that by applying the forecast combination methods such as AI-AFTER, the SA can be (often substantially) improved in most of the considered cases (23 out of 26 cases); this appears to be in alignment with many recent findings that some complicated combination methods on average can significantly outperform simple ones (e.g., Makridakis et al., 2020).

Table 5: Combination forecasts ranked by weighted average losses on last 20 evaluation points.

Forecast	Average Loss (S.E.)
AI-AFTER	0.0337 (0.0034)
CLR	0.0381 (0.0047)
LR	0.0426 (0.0046)
BG_1	0.0448 (0.0043)
SA	0.0503 (0.0049)
AFTER	0.0722 (0.0073)

We then rank the six competing methods by their average losses (across the 26 cases) over the final 20 evaluation points in Table 5, where the different

cases are weighted by the inverse of their full-sample variance (similar to Table VIII of Stock and Watson, 2004b). We see that AI-AFTER and CLR perform the best in this comparison, and AI-AFTER takes the first place that reduces the overall forecast loss of SA by about 33%. The under-performance of AFTER again shows that combining for adaptation is overall not the right goal here; notably, our proposed AI-AFTER significantly improves upon AFTER by intelligently adapting to the proper combining objective to give desirable forecasting performance.

7. Discussion

This work introduces a forecast combining approach, AI-AFTER, that performs well universally in both adaptation and improvement scenarios. By treating methods that attempt to combine for improvement, such as regression-based forecasts, as candidates to be considered and using a hypothesis test to detect underlying forecast scenario, AI-AFTER adapts to the situation at hand to be aggressive or conservative as appropriate based on data and forecast candidates.

So far, our work has focused on the situation where the forecast errors are stationary and the risk is computed under squared error loss. Theoretical and numerical studies of the relative performances of combining for adaptation or improvement under other loss functions, non-stationarity, or in the presence of structural breaks can be of independent interests for future study. In addition, analyzing the theoretical properties of the hypothesis test described in Section 3 to determine the appropriate direction of combining would lead to further understanding and possible refinement of the test.

Supplementary Materials

AIafter The R package implementing our proposed AI-AFTER forecasting algorithm, together with the AFTER, BG, LR, and CLR methods, are available at the GitHub address: `https://github.com/weiqian1/AIafter`.

Acknowledgement

Qian's research is partially supported by NSF grant DMS-1916376 and JPMC Faculty Fellowship. We would like to thank the Editor, Associate Editor and two anonymous referees for their valuable comments that help to improve this manuscript significantly.

References

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Operation Research Quarterly 20, 451–468.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Cheng, X., Hansen, B. E., 2015. Forecasting with factor-augmented regression: A frequentist model averaging approach. Journal of Econometrics 186 (2), 280–293.

Claeskens, G., Magnus, J. R., Vasnev, A. L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32 (3), 754–762.

Clements, M. P., Harvey, D. I., 2011. Combining probability forecasts. International Journal of Forecasting 27 (2), 208–223.

De Luca, G., Magnus, J. R., Peracchi, F., 2018. Weighted-average least squares estimation of generalized linear models. Journal of Econometrics 204 (1), 1–17.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13 (3), 253–263.

Forte, A., Garcia-Donato, G., Steel, M., 2018. Methods and tools for Bayesian variable selection and model averaging in normal linear regression. International Statistical Review 86 (2), 237–258.

Freund, Y., Schapire, R., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (Ed.), Computational Learning Theory. Vol. 904 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 23–37.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164 (1), 130–141.

Granger, C. W. J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3, 197–204.

Granger, C. W. J., White, H., Kamstra, M., 1989. Interval forecasting: an analysis based upon arch-quantile estimators. Journal of Econometrics 40 (1), 87–96.

Hall, S. G., Mitchell, J., 2007. Combining density forecasts. International Journal of Forecasting 23 (1), 1–13.

Hansen, B., 2008. Least-squares forecast averaging. Journal of Econometrics 146, 342–350.

Hendry, D. F., Clements, M. P., 2004. Pooling of forecasts. The Econometrics Journal 7 (1), 1–31.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T., 1999a. Bayesian model averaging: a tutorial. Statistical Science 14, 382–401.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T., 1999b. Bayesian model averaging: A tutorial. Statistical science 14 (4), 382–401.

Hsiao, C., Wan, S. K., 2014. Is there an optimal forecast combination? Journal of Econometrics 178 (2), 294–309.

Ing, C.-K., Lai, T. L., 2011. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. Statistica Sinica 21 (4), 1473–1513.

Kourentzes, N., Barrow, D., Petropoulos, F., 2019. Another look at forecast selection and combination: Evidence from forecast pooling. International Journal of Production Economics 209, 226–235.

Lahiri, K., Peng, H., Sheng, X. S., 2015. Measuring uncertainty of a combined forecast and some tests for forecaster heterogeneity. CESifo Working Paper Series (No. 5468).

Lahiri, K., Peng, H., Zhao, Y., 2017. Online learning and forecast combination in unbalanced panels. Econometric Reviews 36 (1-3), 257–288.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. The M4 competition: Results, findings, conclusion and way forward. International Journal of Forecasting 34 (4), 802–808.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 competition:

100,000 time series and 61 forecasting methods. International Journal of Forecasting 36 (1), 54–74.

Qian, W., Li, W., Sogawa, Y., Fujimaki, R., Yang, X., Liu, J., 2019a. An interactive greedy approach to group sparsity in high dimensions. Technometrics 61 (3), 409–421.

Qian, W., Rolling, C. A., Cheng, G., Yang, Y., 2019b. On the forecast combination puzzle. Econometrics 7 (3), 39.

Shan, K., Yang, Y., 2009. Combining regression quantile estimators. Statistica Sinica 19, 1171–1191.

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

Steel, M. F., 2011. Bayesian model averaging and forecasting. Bulletin of EU and US Inflation and Macroeconomic Analysis 200, 30–41.

Stock, J. H., Watson, M. W., 2003. Forecasting output and inflation: The role of asset prices. Journal of Economic Literature XLI, 788–829.

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

Stock, J. H., Watson, M. W., 2004b. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405–430.

Trapero, J. R., Cardós, M., Kourentzes, N., 2019. Quantile forecast optimal combination to enhance safety stock estimation. International Journal of Forecasting 35 (1), 239–250.

Wallis, K. F., 2005. Combining density and interval forecasts: a modest proposal. Oxford Bulletin of Economics and Statistics 67, 983–994.

Wang, Z., Paterlini, S., Gao, F., Yang, Y., 2014. Adaptive minimax regression estimation over sparse l_q-hulls. The Journal of Machine Learning Research 15 (1), 1675–1711.

Yang, Y., 2003. Regression with multiple candidate models: selecting or mixing? Statistica Sinica 13 (3), 783–809.

Yang, Y., 2004. Combining forecasting procedures: Some theoretical results. Econometric Theory 20 (01), 176–222.

Yang, Y., Qian, W., Zou, H., 2018. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. Journal of Business & Economic Statistics 36 (3), 456–470.

Zhang, T., 2011. Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE Transactions on Information Theory 57 (7), 4689–4708.

Zhang, X., Yu, D., Zou, G., Liang, H., 2016. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. Journal of the American Statistical Association 111 (516), 1775–1790.

Zhang, X., Yu, J., 2018. Spatial weights matrix selection and model averaging for spatial autoregressive models. Journal of Econometrics 203 (1), 1–18.

Zou, H., Yang, Y., 2004. Combining time series models for forecasting. International Journal of Forecasting 20 (1), 69–84.