Combining time series models for forecasting

Hui Zou
Stanford University

Yuhong Yang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/stat_las_preprints

Part of the Statistics and Probability Commons

Recommended Citation
Zou, Hui and Yang, Yuhong, "Combining time series models for forecasting" (2001). Statistics Preprints. 120.
http://lib.dr.iastate.edu/stat_las_preprints/120

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Preprints by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Combining time series models for forecasting

Abstract
Statistical models (e.g., ARIMA models) have commonly been used in time series data analysis and forecasting. Typically, one model is selected based on a selection criterion (e.g., AIC), hypothesis testing, and/or graphical inspection. The selected model is then used to forecast future values. However, model selection is often unstable and may cause an unnecessarily high variability in the final estimation/prediction. In this work, we propose the use of an algorithm, AFTER, to convexly combine the models for a better performance of prediction. The weights are sequentially updated after each additional observation. Simulations and real data examples are used to compare the performance of our approach with model selection methods. The results show an advantage of combining by AFTER over selection in terms of forecasting accuracy at several settings.

Keywords
ARIMA, combining forecasts, forecast instability, model selection, modeling

Disciplines
Statistics and Probability

Comments
Combining Time Series Models for Forecasting

Yuhong Yang  
Department of Statistics  
Snedecor Hall  
Iowa State University  
Ames, IA 50011-1210  
Phone: 515-294-2089  
Fax: 515-294-4040  
Email: yyang@iastate.edu

Hui Zou  
Department of Statistics  
Sequoia Hall  
Stanford University  
Stanford, CA 94305-4065  
Phone: 650-725-5952  
Fax: 650-725-8977  
Email: hzou@Stanford.edu

October, 2001
Abstract

Statistical models (e.g., ARIMA models) have been commonly used in time series data analysis and forecasting. Typically one model is selected based on a selection criterion (e.g., AIC), hypothesis testing, and/or graphical inspections. The selected model is then used to forecast future values. However, model selection is often unstable and may cause an unnecessarily high variability in the final estimation/prediction. In this work, we propose the use of an algorithm AFTER to convexly combine the models for a better performance of prediction. The weights are sequentially updated after each additional observation.

Simulations and real data examples are used to compare performance of our approach with model selection methods. The results show advantage of combining by AFTER over selection in term of forecasting accuracy at several settings.

Keywords: Combining forecasts, forecast instability, ARIMA modeling, model selection.


Hui Zou received his BS and MS degree in physics from the University of Science and Technology of China in 1997 and 1999 respectively. He recently graduated from the Department of Statistics at Iowa State University with MS degree. He is now a Ph.D. student in the Department of Statistics at Stanford University. He has several publications in physics journals including International Journal of Modern Physics A, Physics Letters A, Modern Physics Letters A, and Modern Physics Letter B.
1 Introduction

Let $Y_1, Y_2, \ldots$ be a time series. At time $n$ for $n \geq 1$, we are interested in forecasting or predicting the next value $Y_{n+1}$ based on the observed realizations of $Y_1, \ldots, Y_n$. We focus on one-step ahead point forecasting in this work.

Statistical models have been widely used for time series data analysis and forecasting. For example, ARIMA modeling approach proposed by Box and Jenkins (1976) has been proven to be effective in many applications relative to ad hoc forecasting procedures. In a practical situation, for applying the statistical modeling approach, however, one faces the important issue of how to choose the “best” model among a variety of candidates. Generally speaking, the issue is highly non-trivial and has received considerable attention with different approaches being proposed and studied. We briefly discuss some of these approaches that are closely related to our work. Readers are referred to de Gooijer, et al (1985) for a review on this topic.

Graphical inspection together with examination of simple summary statistics (such as autocorrelations) is very useful and should be used, by all means, for preliminary analysis. However, for the purpose of selecting a model, this approach is ad hoc and too subjective in general. The use of statistical hypothesis testing techniques is more formal in nature. However, there are difficulties with this approach as well. Firstly one faces the challenging issue of multiple testing, and due to the sequential nature of the tests, there is little one can say about the probabilities of errors associated with the whole procedure after conducting a series of tests. Secondly, there is no objective guideline for the choice of the size of each individual test and it is completely unclear how such a choice affects the forecasting accuracy. In addition, even when one compares only two models, the model preferred by a test (or even the true model) does not necessarily perform better than the other one in terms of prediction risk. Another approach uses a well defined and formal model selection criterion. Model selection criteria have been proposed based on different considerations, e.g., AIC (Akaike (1973)) by considering a discrepancy measure between the true model and a candidate, and BIC (Schwarz (1978)) by considering approximating the posterior model probabilities in a Bayesian framework. Hannan and Quinn (1979) proposed a related criterion for AR models which has a smaller penalty compared to BIC yet still permits a strong consistency property. The approach of using a model selection criterion has now been commonly utilized in practice. It is interesting that Chatfield (2001, p. 47) noted that econometricians tend to follow the hypothesis testing approach while statisticians tend to use model selection methods (of course with appropriate model checking after selection).

One major drawback with model selection is its instability. With a small or moderate number of observations, as expected, models close to each other are usually hard to be distinguished and the model selection criterion values are usually quite close to each other. The choice of the model with the smallest criterion value is accordingly unstable for such a case. A slight change of the data may result in the choice of a different model. As a consequence, the forecast based on the selected model may have a high
variability.

In this work, we propose the use of a method, named AFTER, to combine the forecasts from the individual candidate models. The hope is that with an appropriate weighting scheme, the combined forecast has a smaller variability so that the forecasting accuracy can be improved relative to the use of a selection criterion. We will focus on the ARIMA models, and simulations and some real data sets will be used to compare combining and selection in terms of forecasting accuracy. As will be seen, combining tends to reduce the prediction error when there is difficulty in identifying the best model.

Combining forecasts has been studied for the past three decades (see Clemen (1989) for a comprehensive review of this topic). Various methods have been proposed. The focus has been on the case when the forecasts to be combined are distinct in nature (i.e., based on very different methods). For example, Clements and Hendry (1998, Chapter 10) stated that “When forecasts are all based on econometric models, each of which has access to the same information set, then combining the resulting forecasts will rarely be a good idea. It is better to sort out the individual models - to derive a preferred model that contains the useful features of the original models”. In response to the criticisms of the idea of combining, Newbold and Granger (1974) wrote “If (the critics) are saying ... that combination is not a valid proposition if one of the individual forecasts does not differ significantly from the optimum, we must of course agree”. In our view, however, combining (mixing) forecasts from very similar models is also important. For the reason mentioned earlier, combining has a great potential to reduce variability that arises in the forced action of selecting a single model. The empirical results in this paper do support our point of view.

Instability of model (or procedure) selection has been recognized in statistics and related literature (e.g., Breiman (1996)). When multiple models are considered for estimation and prediction, the term “model uncertainty” is used to capture the difficulty in identifying the correct model by several authors (e.g., Chatfield (1996) and Hoeting et al. (1999)). In our point of view, the term makes most sense when it is interpreted as the uncertainty in finding the “best” model. Here best may be defined in terms of an appropriate loss function (e.g., square error loss in prediction). We take the point of view that in general the “true” model may or may not be in the candidate list and even if the true model happens to be included, the task of finding the true model can be very different from that of finding the best model for the purpose of prediction.

We are not against the practice of model selection in general. Identifying the true model (when it makes good sense) is an important task to understand relationships between variables. In linear regression, it is observed that selection may outperform combining methods when one model is very strongly preferred, in which case there is little instability in selection. In the time series context, our observation is that again when model selection is stable, combining does not necessarily lead to improvement.

We should also point out that the approach of combining we take is related but different from formal Bayesian consideration. Particularly, no prior distributions will be considered for parameters in the models.
The rest of the paper is organized as follows. In Section 2, we briefly review some model selection criteria and address some basic issues. In Section 3, the combining algorithm AFTER is presented. In Section 4, we give results of several simulations studies for comparing combining and selection. Examples of real data are used in Section 5 to demonstrate an advantage of combining by AFTER. Conclusions and discussions are in Section 6.

2 Some preliminaries

2.1 Evaluation of forecasting accuracy

Assume that for \( i \geq 1 \), the conditional distribution of \( Y_i \) given the previous observations \( Y^{i-1} = \{Y_j \}_{j=1}^{i-1} \) has (conditional) mean \( m_i \) and variance \( v_i \). That is, \( Y_i = m_i + \epsilon_i \), where \( \epsilon_i \) is the random error that represents the conditional uncertainty in \( Y \) at time \( i \). Note that \( E(\epsilon_i | Y^{i-1}) = 0 \) almost surely for \( i \geq 1 \).

Let \( \hat{Y}_i \) be a predicted value of \( Y_i \) based on \( Y^{i-1} \). Then the one-step ahead mean square prediction error is

\[
E \left( Y_i - \hat{Y}_i \right)^2.
\]

Very naturally it can be used as a performance measure for forecasting \( Y_i \). Note that the conditional one-step ahead forecasting mean square error can be decomposed into squared bias and conditional variance as follows:

\[
E_i \left( Y_i - \hat{Y}_i \right)^2 = \left( m_i - \hat{Y}_i \right)^2 + v_i^2,
\]

where \( E_i \) denotes the conditional expectation given \( Y^{i-1} \). The latter part is not in one’s control and is always present regardless of which method is used for prediction. Since \( v_i \) is the same for all forecasts, it is also sensible to remove it in measuring performance of a forecasting procedure. Accordingly, for forecasting \( Y_i \), we may consider the loss function

\[
L(Y_i, \hat{Y}_i) = \left( m_i - \hat{Y}_i \right)^2
\]

and the corresponding risk is

\[
E \left( m_i - \hat{Y}_i \right)^2.
\]

Let \( \delta \) be a forecasting procedure that yields forecasts \( \hat{Y}_1, \hat{Y}_2, ... \) at times 1, 2 and so on. Two performance measures are natural for consideration to compare different forecasting procedures. The first is the (sequential) average mean square error in prediction from forecasting \( Y_{n+1} \) up to predicting \( Y_{n+1} \), denoted by \( AMSEP(\delta; n_0; n) \), as follows

\[
AMSEP(\delta; n_0; n) = \frac{1}{n - n_0 + 1} \sum_{i=n_0+1}^{n+1} E \left( Y_i - \hat{Y}_i \right)^2.
\]

The observations \( \{Y_i\}_{i=1}^{n_0} \) are used for initial estimation and for each \( i = n_0 + 1, ..., n+1 \), forecasts are obtained with all the available observations. For the particular case of \( n_0 = n \), it is simply the mean
square error in prediction at time $n$, and will be denoted $MSEP(\delta; n)$. The second performance measure is the average net mean square error in prediction

$$ANMSEP(\delta; n_0; n) = \frac{1}{n - n_0 + 1} \sum_{i=n_0+1}^{n+1} E \left( m_i - \hat{Y}_i \right)^2.$$  

For the case of $n_0 = n$, it is the net mean square error in prediction at time $n$, and will be denoted $NMSEP(\delta; n)$. We will focus on ANMSEP for comparing model selection with model combining in simulations.

For the purpose of comparing forecasting procedures based on real data, given a time series of size $n$, we will consider the (sequential) average square error in prediction

$$ASEP(\delta; n_0; n) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \left( Y_i - \hat{Y}_i \right)^2$$

as a performance measure of a forecasting procedure with $n_0$ being a fraction (but not too small) of $n$. Clearly unlike the theoretical quantities $AMSEP$ and $ANMSEP$, it can be computed based on the data alone.

### 2.2 Some model selection criteria

Let $B$ denote the backward shift operator, i.e., $BY_i = Y_{i-1}$. ARMA models take the form

$$\phi(B)Y_i = \theta(B)\varepsilon_i,$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are polynomials of finite orders $p$ and $q$, respectively. If the $d$-th difference of $\{Y_i\}$ is an ARMA process of order $p$ and $q$, then $Y_i$ is called an ARIMA($p,d,q$) process. ARIMA modeling has now become a standard practice in time series analysis. It is a practically important issue to determine the orders $p$, $d$, and $q$.

For a given set of $(p,d,q)$, one can estimate the unknown parameters in the model by e.g., maximum likelihood method. AIC (Akaike 1973) selects the model that minimizes the criterion

$$-\log(\text{maximized likelihood}) + p + q.$$  

BIC (Schwarz 1978) selects the model that minimizes

$$-\log(\text{maximized likelihood}) + (p + q) \cdot \ln n/2.$$  

Hannan and Quinn (1978) proposed to minimize

$$-\log(\text{maximized likelihood}) + (p + q) \cdot \ln \ln n/2.$$  

Small sample correction of AIC has been studied by e.g., Hurvich and Tsai (1989). These criteria will be considered in the empirical studies later in this paper.

Theoretical properties of these criteria have been investigated. It is known that BIC and HQ are consistent in the sense that the probability of selecting the true model approaches 1 (if the true model
is in the candidate list) but AIC is not (see e.g., Shibata (1976) and Hannan (1982)). On the other hand, Shibata (1980) showed that AIC is asymptotically efficient in the sense that the selected model performs (under the mean square error) asymptotically as well as the best model when the true model is not in the candidate list. Due to the asymptotic nature, these results provide little guidance for real applications with a small or moderate number of observations.

Common to all model selection methods is the potentially large variability in the resulting estimator/forecast. Proper weighting of the competing models may substantially improve the performance. In a simplified density estimation context, Yang (2001c) showed an advantage of a proper combining over any selection method.

### 2.3 Identifying the true model is not necessarily optimal for forecasting

Suppose that two models are being considered for fitting a time series data. When the two models are nested, very naturally, one can employ a hypothesis testing technique (e.g., likelihood ratio test) to assess which of them is more likely to be the one that generated the data (assuming that at least one of the models is correct). Since tests of an exact given size are often hard to find or compute for time series data, asymptotic tests might be used instead. For a small or moderate number of observations, the asymptotic approximation may not be accurate enough. Furthermore, a more serious concern with the approach of attempting to identify the true model for forecasting is that even if one correctly identified the true model, the forecast based on the model actually may not perform as well as a wrong model in terms of forecasting accuracy. As is well-known in regression, due to the trade-off between bias and variance, performance of the true model may be worse than a simpler model. Here we give a simple example for an illustration and the example will be revisited to understand difference between combining and selection.

**Example 0:** Consider two simple models: for $n \geq 1$,

\[
Y_n = \epsilon_n,
\]

and

\[
Y_n = \alpha Y_{n-1} + \epsilon_n,
\]

where $\{\epsilon_t\}$ are i.i.d. normally distributed with mean zero and unknown variance $\sigma^2$. Here we assume that $0 \leq \alpha < 1$. Obviously, model 1 is nested in model 2 with $\alpha = 0$.

Under model 1, the mean square prediction risk is clearly minimized when $\hat{Y}_{n+1} = 0$. If actually $\alpha$ is nonzero, then the MSEP is

\[
E \left( \hat{Y}_{n+1} - Y_{n+1} \right)^2 = \sigma^2 + \alpha^2 EY_n^2.
\]

For model 2, with $\alpha$ estimated by $\hat{\alpha}$, a natural forecast is $\hat{Y}_{n+1} = \hat{\alpha} Y_n$ and then the MSEP is

\[
E \left( \hat{Y}_{n+1} - Y_{n+1} \right)^2 = \sigma^2 + E(\hat{\alpha} - \alpha)^2 Y_n^2.
\]
For comparing performance of the two models, we can examine either the difference

$$a^2 EY_n^2 - E (\hat{a}_n - a)^2 Y_n^2$$

or the ratio

$$\frac{a^2 EY_n^2}{E (\hat{a}_n - a)^2 Y_n^2}.$$  

(1)

It does not seem to be easy to examine the two quantities analytically. Monte Carlo simulations can be used instead.

Figure 1 gives the graphs of the difference/ratio defined above in $a$, with $n = 20$ and $\sigma^2$ fixed to be 1 based on Monte Carlo simulations with 100 replications for each choice of $a$. The upper two graphs plot the risk difference against $a$ with the second one focused on $a$ in the range from 0 to 0.5. The lower two graphs treat the ratio in (1) instead of the difference.

From the graphs, clearly, when $a$ is small (less than about 0.3), the wrong model performs better. When $a$ gets larger, however, the true model works better. It is unclear how the significance level of a test relates to which model is better in prediction.

This example was considered by Chatfield (2001, pp. 220-221). He investigated the effect on parameter estimation of $a$ based on a test (in terms of the first order auto-correlation coefficient) to decide whether $a$ is different from zero. He found through simulations that substantial bias is introduced this way.
2.4 Measuring stability of model selection methods

As mentioned earlier (see also Section 5), model selection can be very unstable. When there is a substantial uncertainty in finding the best model, alternative methods such as model combining should be considered. An interesting issue then is: How to measure stability of model selection or the resulting prediction? We here propose two simple approaches.

2.4.1 Sequential stability

Consider a model selection method. One idea of measuring stability in selection is to examine its consistency in selection at different data sizes. Suppose that the model \( \hat{k}_n \) is selected by the method based on all the observations \( \{Y_i\}_{i=1}^n \). Let \( L \) be an integer between 1 and \( n - 1 \). For each \( j \) in \( \{n - L, n - L + 1, \ldots, n - 1\} \), apply the model selection method to the data \( \{Y_i\}_{i=1}^j \) and let \( \hat{k}_j \) denote the selected model. Then let \( \kappa \) be the percentage of times that the same model \( (\hat{k}_n) \) is selected, i.e.,

\[
\kappa = \frac{\sum_{j=n-L}^{n-1} I_{\{\hat{k}_j = \hat{k}_n\}}}{L},
\]

where \( I_{\{\hat{k}_j = \hat{k}_n\}} \) denotes the indicator function. The rationale behind the consideration of \( \kappa \) is quite clear: removing a few observations should not cause much change for a stable procedure. The integer \( L \) should be chosen appropriately. On one hand, one wants to choose \( L \) small so that the selection problems for \( j \) in \( \{n - L, \ldots, n - 1\} \) are similar to the real problem (with the full data observed). On the other hand, one needs to have \( L \) not too small so that \( \kappa \) is reasonably stable.

It should be pointed out that the measure here does not address directly the issue of selecting the best model. A model selection method (e.g., the trivial one that always selects the largest model) may well be stable but perform poorly for forecasting. If a method is unstable for the data in the sequential stability, its ability to pick up the best model is in doubt.

2.4.2 Perturbation stability

Another approach to measuring stability in model selection is through perturbation. The idea is simple: if a statistical procedure is stable, a minor perturbation of the data should not change the outcome dramatically.

Consider a model selection criterion for comparing ARMA models

\[
\phi(B)Y_i = \theta(B)\epsilon_i,
\]

where \( \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \) and \( \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q \). Let \( \hat{p} \) and \( \hat{q} \) be the orders selected by the criterion. Let \( \hat{\phi}(B) = 1 - \hat{\phi}_1 B - \cdots - \hat{\phi}_p B^p \) and \( \hat{\theta}(B) = 1 + \hat{\theta}_1 B + \cdots + \hat{\theta}_q B^q \), where \( \hat{\phi}_i \) and \( \hat{\theta}_i \) are parameter estimates based on the data. Now we generate a time series following the model

\[
\hat{\phi}(B)W_i = \hat{\theta}(B)\eta_i,
\]

where \( \eta_i \) are i.i.d. \( \sim N(0, \tau^2\hat{\sigma}^2) \) with \( \tau > 0 \) and \( \hat{\sigma}^2 \) being an estimate of \( \sigma^2 \) based on the selected model. Consider now \( \bar{Y}_i = Y_i + W_i \) for \( 1 \leq i \leq n \) and apply the model selection criterion to the new data \( \{\bar{Y}_i\}_{i=1}^n \).
If the model selection criterion is stable for the data, then when \( \tau \) is small, the newly selected model is most likely the same as before and the corresponding forecast should not change too much either. For each \( \tau \), we generate \( \{W_t\}_{t=1}^{\infty} \) a large number of times (say 100) independently.

**Stability in selection** For each \( \tau \), we record the percentage of times that the originally selected model is chosen again with the perturbation. We then plot the percentage versus \( \tau \). If the percentage decreases sharply in \( \tau \), it indicates that the selection procedure is unstable even with small perturbation.

**Instability in forecasting** Stability in selection does not necessarily capture the stability in forecasting, because different models may perform equally well in prediction. Alternatively, at each \( \tau \), we compute the average deviation of the new forecast from the original one relative to the estimated \( \sigma \) (using the initially selected model), based on a large number, say 100, of replications. That is, we average

\[
\frac{|\hat{y}_{n+1} - \hat{y}_{n+1}|}{\sigma}
\]

over 100 independent perturbations at size \( \tau \), where \( \hat{y}_{n+1} \) is obtained by applying the selection procedure again on the perturbed data, and \( \hat{y}_{n+1} \) and \( \sigma \) are based on the original data. It will be called the forecast perturbation instability of the procedure at perturbation size \( \tau \) for the given data. Again, how fast this quantity increases in \( \tau \) is a reasonable instability measure.

### 2.4.3 Data examples

Consider the data sets 1 and 4 (See Section 5 for details).

For data set 1, AR(1) fits very well. Figure 1 gives the perturbation stability plots over AR models with order up to 5. There is little uncertainty in model selection (AIC, BIC and HQ all select AR(1) and the sequential stability \( \kappa \) is 1 for the selection methods for \( L \) not close to \( n \)). From the graph, clearly, a small perturbation does not really change the outcome of selection and changes very little the forecast.

For data set 4, a log transformation is used. ARIMA\((p,d,q)\) models with \( p = 0, ..., 5, d = 0, 1, q = 0, ..., 5 \) are the candidates. Figure 3 is the forecast instability plots of the model selection methods. For each of AIC, BIC and HQ, the stability in selection drops to near zero very sharply in \( \tau \).

Compared to data set 1, the perturbation forecast instability of data set 4 is dramatically higher for the model selection methods, especially for BIC and HQ. This suggests that the model selection criteria have difficulty finding the best model. As will be seen later in Section 5, AFTER can improve forecasting accuracy for this case, while there is no advantage in combining for data set 1. The perturbation stability plots can be useful for deciding whether to combine or select in a practical situation.

In addition to the above approaches, other methods based on resampling (e.g., bootstrapping) are possible to measure model instability. We will not pursue those directions in this work.
3 Algorithm AFTER for combining forecasts

Assume that the conditional distribution of $Y_i$ given $Y^{i-1} = y^{i-1}$ is Gaussian for all $i \geq 1$ with conditional mean $m_i$ and the conditional variances $v_i$. Assume that for each forecasting procedure $\delta_j$, in addition to the forecast $\hat{y}_n$, an estimate of $v_n$, say, $\hat{v}_{jn}$ is obtained based on $y^{n-1}$. If the observations are stationary, various variance estimation methods have been proposed for different scenarios. Note that the procedures do not have to use different variance estimators and if some procedures do not provide variance estimates, we can borrow from others.

Yang (2001b) proposed an algorithm AFTER to combine different forecasts. He examined its theoretical convergence properties. In this work, we apply AFTER to the case that multiple models of the same type are considered for forecasting.

To combine the forecasting procedures $\Delta = \{\delta_1, \delta_2, ..., \delta_J\}$, at each time $n$, the AFTER algorithm looks at their past performances and assign weights accordingly as follows.

Let $W_{j,1} = 1/J$ and for $n \geq 2$, let

$$W_{j,n} = \frac{\prod_{i=1}^{n-1} \hat{v}_{ji,i}^{-1/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{n-1} \frac{(y_i - \hat{y}_{ji,i})^2}{\hat{v}_{ji,i}} \right)}{\sum_{j' \geq 1} \prod_{i=1}^{n-1} \hat{v}_{j'i,i}^{-1/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{n-1} \frac{(y_i - \hat{y}_{j'i,i})^2}{\hat{v}_{j'i,i}} \right)}. \quad (2)$$
Figure 3. Forecast Instability for Data Set 4

Then combine the forecasts by

\[ \hat{y}_n^* = \sum_{j=1}^{J} W_{j,n} \hat{y}_{j,n}. \]

Note that \( \sum_{j=1}^{J} W_{j,n} = 1 \) for \( n \geq 1 \) (thus the combined forecast is a convex combination of the original ones), and \( W_{j,n} \) depends only on the past forecasts and the past realizations of \( Y \).

Note also that

\[ W_{j,n} = \frac{W_{j,n-1} \hat{v}_{j,n-1}^{1/2} \exp \left( \frac{-(Y_{n-1} - \hat{y}_{j,n-1})^2}{2\hat{v}_{j,n-1}} \right)}{\sum_{j'} W_{j',n-1} \hat{v}_{j',n-1}^{1/2} \exp \left( \frac{-(Y_{n-1} - \hat{y}_{j',n-1})^2}{2\hat{v}_{j',n-1}} \right)}. \] (3)

Thus after each additional observation, the weights on the candidate forecasts are updated. The algorithm is called Aggregated Forecast Through Exponential Re-weighting (AFTER).

Remarks:

1. Under some conditions, Yang (2001b) shows that the risk of the combined procedure satisfies

\[
\frac{1}{n} \sum_{i=1}^{n} E \left( \frac{(m_i - \hat{Y}_i^*)^2}{v_i} \right) \leq c \inf_{j \geq 1} \left( \log J + \frac{1}{n} \sum_{i=1}^{n} E \left( \frac{(m_i - \hat{Y}_{i,j})^2}{v_i} \right) + \frac{1}{n} \sum_{i=1}^{n} E \left( \frac{\hat{v}_{j,i} - v_i^2}{v_i^2} \right) \right),
\]

where \( c \) is a reasonable constant. Thus the combined forecasting procedure is automatically within a multiple of the best one plus the risk of variance estimation. Consequently, with appropriate variance estimation, AFTER yields automatically the best rate of convergence provided by the individual forecasting procedures. It is worth noting that the result does not require any of the forecasting procedure be based on the “true” model. Similar results in the context of regression are in, e.g., Yang (2001a).
2. The weighting in (3) has a Bayesian interpretation. If we view \( W_{j,n-1}, j \geq 1 \) as the prior probabilities on the procedures before observing \( Y_{n-1} \), then \( W_{j,n} \) is the posterior probability of \( \delta_j \) after \( Y_{n-1} \) is seen. However, our procedure is not a formal Bayesian one. No prior probability distributions are considered for the parameters.

3. The normality assumption is not essential. As long as the distribution of the error is known up to a scale parameter, similar combining methods are given in Yang (2001a).

4 Simulations

4.1 Two simple models

This subsection continues from Section 2.3. We intend to study differences between selection and combining in that simple setting.

Consider risks of forecasts based on selection and combining. Fix a model selection method. Let \( A \) denote the set of sample points on which model 1 is accepted by the method. Then the associated risk is

\[
E \left( \left( \bar{Y}_{n+1} - Y_{n+1} \right)^2 I_A + E \left( Y_{n+1}^2 \right) I_{A^c} \right),
\]

where \( I \) is the indicator and \( A^c \) denotes the complement of \( A \). For a combining method, let \( W_1 \) and \( W_2 \) be the (non-negative) weights on model 1 and 2 respectively. Note that \( W_1 \) and \( W_2 \) depend on the observations and add up to 1 for convex combining. The corresponding risk is

\[
E \left( W_1 \left( \bar{Y}_{n+1} - Y_{n+1} \right)^2 + E \left( W_2 Y_{n+1}^2 \right) \right).
\]

The following is a result from a simulation study with \( n = 20 \) based on 1000 replications. For applying the AFTER method, we start with the first 10 observations with equal initial weights on the two models. Table 1 gives percentages of selecting the true model by AIC, BIC, HQ and AICc (Hurvich and Tsai (1989)). Table 2 presents the risks of the true model, wrong model, AIC, BIC, HQ and AFTER as defined earlier at three representative \( \alpha \) values.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.157</td>
<td>0.094</td>
<td>0.135</td>
<td>0.144</td>
</tr>
<tr>
<td>0.50</td>
<td>0.580</td>
<td>0.463</td>
<td>0.557</td>
<td>0.568</td>
</tr>
<tr>
<td>0.91</td>
<td>0.931</td>
<td>0.884</td>
<td>0.921</td>
<td>0.924</td>
</tr>
</tbody>
</table>

Table 1: Percentage of Selecting the True Model

Figures 4 and 5 compare the model selection criteria with AFTER in terms of \( ANMSEP(\delta; 10; 20) \) and \( NMSEP(\delta; 20) \) respectively. Note that even with 1000 replications, the simulated risks for the model selection methods still fluctuate significantly, which indicates instability of model selection.

It seems that when \( \alpha \) is close to the extreme, the selection criteria choose the best model (not necessarily the true model) with very high probability, i.e., there is little instability in selection. Then AIC, BIC and HQ perform quite well.
It is clearly seen that as \( \alpha \) increases, AFTER puts higher weight on the right model as is desired.

Table 2: Comparing Risks of Model Selection and Combining for the Two Model Case

<table>
<thead>
<tr>
<th>( \alpha = 0.01 )</th>
<th>True</th>
<th>Wrong</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AFTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.100 (0.005)</td>
<td>0.000 (0.000)</td>
<td>0.029 (0.004)</td>
<td>0.016 (0.003)</td>
<td>0.024 (0.003)</td>
<td>0.029 (0.003)</td>
<td></td>
</tr>
<tr>
<td>0.122 (0.007)</td>
<td>0.340 (0.015)</td>
<td>0.165 (0.010)</td>
<td>0.204 (0.012)</td>
<td>0.168 (0.010)</td>
<td>0.146 (0.008)</td>
<td></td>
</tr>
<tr>
<td>0.340 (0.017)</td>
<td>4.536 (0.217)</td>
<td>0.333 (0.060)</td>
<td>0.375 (0.061)</td>
<td>0.350 (0.060)</td>
<td>0.440 (0.028)</td>
<td></td>
</tr>
</tbody>
</table>

It is interesting to note that when the average risk \( ANMSEP(\cdot;10,20) \) is considered, the advantage of AFTER is stronger, because smaller sample sizes are involved in \( ANMSEP(\cdot;10,20) \), where instability in selection is more severe.

Table 3 gives the average weights that AFTER put on the two models at the end of 20 observations. It is clearly seen that as \( \alpha \) increases, AFTER puts higher weight on the right model as is desired.

4.2 AR models with different orders

The candidate models are AR models with orders up to 8 unless stated otherwise. We consider the risk \( ANMSEP(\cdot; n_0; n) \) with different choices of \( n_0 \) and \( n \). The error variance \( \sigma^2 \) is fixed to be 1. Since the first \( n_0 \) observations are available in the beginning (the competition begins with forecasting \( Y_{n_0+1} \) till time \( n \), for AFTER, instead of using the uniform weights on the models for predicting \( Y_{n_0+1} \), we can
begin weighting based on the first $m$ observations with $m$ smaller than $n_0$ (but not too small to avoid inaccurate estimation due to too small sample size). This way, the weights at $n_0$ will be more suitable than equal weighting. When $n - n_0$ is large, the difference is small.

4.2.1 Case 1

The true model is AR(1) with coefficient 0.8. The candidates are AR models up to order 5. We take $n = 50$ and $n_0 = 15$ or 50. For AFTER, weighting starts at 5. The results are in Table 4 based on 200 replications.

<table>
<thead>
<tr>
<th>$n_0$</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AFTER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_0 = 15$</td>
<td>$n_0 = 50$</td>
<td>$n_0 = 15$</td>
<td>$n_0 = 50$</td>
</tr>
<tr>
<td></td>
<td>0.120</td>
<td>0.095</td>
<td>0.102</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.010)</td>
<td>(0.007)</td>
<td>(0.009)</td>
</tr>
<tr>
<td></td>
<td>0.109</td>
<td>0.085</td>
<td>0.109</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.010)</td>
<td>(0.007)</td>
<td>(0.009)</td>
</tr>
<tr>
<td></td>
<td>0.119</td>
<td>0.093</td>
<td>0.096</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.009)</td>
<td>(0.006)</td>
<td>(0.009)</td>
</tr>
</tbody>
</table>

Table 4: Comparing Model Selection to Combining with AR models: Case 1
For this case, the model selection criteria have little difficulty finding the best model and AFTER is not expected to be advantageous. In fact, it performs significantly worse than BIC at both \( n_0 = 15 \) and \( n_0 = 50 \).

### 4.2.2 Case 2

Here the true model is AR(2) with coefficients \((0.5864, -0.15)\). The results for \( n = 50 \) and \( n_0 = 20 \) or 50 are in Table 5 based on 500 replications. For AFTER, weighting starts at 15.

<table>
<thead>
<tr>
<th>( n_0 = 20 )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AFTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.104</td>
<td>0.092</td>
<td>0.010</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td></td>
</tr>
<tr>
<td>( n_0 = 50 )</td>
<td>0.079</td>
<td>0.073</td>
<td>0.075</td>
<td>0.064</td>
</tr>
<tr>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.005)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Comparing Model Selection to Combining with AR models: Case 2

### 4.2.3 Case 3

Here the true model is AR(3) with coefficients \((0.7, -0.3, 0.2)\). The results for \( n = 50 \) and \( n_0 = 15 \) or 50 are in Table 6 based on 200 replications. For AFTER, weighting starts at 15.

<table>
<thead>
<tr>
<th>( n_0 = 15 )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AFTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.155</td>
<td>0.159</td>
<td>0.156</td>
<td>0.126</td>
<td></td>
</tr>
<tr>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td></td>
</tr>
<tr>
<td>( n_0 = 50 )</td>
<td>0.148</td>
<td>0.143</td>
<td>0.156</td>
<td>0.114</td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.012)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Comparing Model Selection to Combining with AR models: Case 3

### 4.2.4 Case 4

Here the true model is AR(4) with coefficients \((0.9, -0.5, 0.2, 0.2)\). We take \( n = 100 \), \( n_0 = 50 \) or 100 based on 200 replications. After begins weighting at 50. The results are given in Table 7.

<table>
<thead>
<tr>
<th>( n_0 = 50 )</th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
<th>AFTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.132</td>
<td>0.160</td>
<td>0.137</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td>(0.005)</td>
<td></td>
</tr>
<tr>
<td>( n_0 = 100 )</td>
<td>0.097</td>
<td>0.100</td>
<td>0.090</td>
<td>0.082</td>
</tr>
<tr>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.008)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Comparing Model Selection to Combining with AR models: Case 4

### 4.2.5 Random models

To show that the advantage of AFTER seen above are not atypical, we compare AFTER with model selection in a random setting. We randomly select the true AR order (uniformly between 1 and 8) and
then randomly generate the coefficients with uniform distribution on $[-10, 10]$ (discard the case if the coefficients do not yield stationarity). One hundred ten true models were obtained this way. Table 8 compares AFTER with AIC, BIC, and HQ by examining the risk difference and the ratio of the net risks for $n = 100$ and $n_0 = 50$ based on 100 replications. For AFTER, weighting begins at 50.

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>HQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Difference</td>
<td>0.036</td>
<td>0.044</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.006)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Risk Ratio</td>
<td>1.063</td>
<td>1.678</td>
<td>1.578</td>
</tr>
<tr>
<td></td>
<td>(0.092)</td>
<td>(0.133)</td>
<td>(0.099)</td>
</tr>
</tbody>
</table>

Table 8: Comparing Model Selection to Combining with AR models: Random Case

The table clearly shows that AFTER performs significantly better in prediction. Figures 6 and 7 are the box-plots of the risks of AIC, BIC, HQ and AFTER, and the risks of AIC, BIC, and HQ relative to that of AFTER.

![Box-plot of risks](image)

**Figure 6.** Risks of AIC, BIC, HQ and AFTER with Random AR Models

Note that based on the 110 replications, AFTER has a smaller risk compared to AIC, BIC, and HQ about 98%, 89%, and 91% of times respectively.

### 4.3 ARIMA models with different orders

Here we compare the model selection methods with AFTER based on ARIMA models with $n = 100$ and $n_0 = 50$ with 100 replications. The true model is ARIMA$(2, 1, 2)$ with AR coefficients $(0.8, 0.1)$ and MA coefficients $(0.6, 0.1)$. The candidate models are ARIMA$(p, d, q)$ with $p, q = 1, \ldots, 8$, $d = 0, 1$. Figure 8 shows the box-plots of the differences of the net square error losses of AIC, BIC and HQ relative to that of AFTER. Clearly AFTER performs better averagely speaking. In fact, based on the simulation,
AFTER has a smaller loss 80%, 82%, and 78% of times compared to AIC, BIC and HQ, respectively.

5 Data examples

We compare AFTER with model selection on real data. We use $ASEP(\delta; n_0; n)$ as performance measure. Obviously, $n_0$ should not be too close to $n$ (so that the ASEP is reasonably stable) but not too small (in which case there are too few observations to build models with reasonable accuracy in the beginning). For applying AFTER, we begin weighting based on the first $m$ observations with $m$ being smaller than $n_0$ by 5 or 10.
5.1 Data set 1

The observations were the daily average number of defects per truck at the end of the assembly line in a manufacturing plant for \( n = 45 \) consecutive business days (see, e.g., Wei (1990, p. 446)). Wei suggests AR(1) model for the data and it indeed fits very well. Take \( n_0 = 30 \) and consider model selection methods AIC, BIC and HQ as well as AFTER (weighting starts at 25). Then the \( ASEP(n_0; n) \) for AIC, BIC, and HQ are all equal to 0.250 and is 0.254 for AFTER. The sequential stability \( \kappa (L = 15) \) is 1 for all the selection methods and they all choose AR(1). This is a case where there is little instability in model selection and AFTER has no advantage.

5.2 Data set 2

This data set consists of levels of Lake Huron in feet for July from 1875 to 1972 (see, e.g., Brockwell and Davis (1991, p. 328)) with \( n = 98 \). Graphic inspections suggested ARMA(1, 1) for the data and the residual plots look very reasonable.

Here we consider candidates models ARIMA\((p, d, q)\), \( p = 0, 1, 2; d = 0, 1; q = 0, 1, 2 \). For \( n_0 = 78 \), the \( ASEP(n_0; n) \) for AIC, BIC, and HQ are 0.778, 0.714, and 0.714 respectively. For AFTER with weighting beginning at 68, the \( ASEP(n_0; n) \) is 0.668, a reduction of at least 6%. The sequential stability \( \kappa \) (with \( L = 20 \)) is 0.50, 1, and 1 for AIC, BIC and HQ, respectively. Note that for the different sample sizes between 78 and 98, BIC and HQ always selected ARMA(1,1), but AIC tended to select ARMA(2,1). Somehow BIC and HQ are more stable for this data than AIC.

5.3 Data set 3

The data are monthly numbers of unemployed females between Ages 16 and 19 in the United states from January 1961 to December 1985 (in thousands) (see, e.g., Wei (1990, p. 448)) with a total of \( n = 300 \) observations.

We consider ARIMA\((p, d, q)\) models with \( p = 0, 1, 2; d = 0, 1; q = 0, 1, 2 \). For \( n_0 = n = 20 \), the \( ASEP(n_0; n) \) for AIC, BIC, HQ, and AFTER (with weighting beginning at \( n - 30 \)) are 1739, 1731, 1731, and 1672, respectively. The sequential stability \( \kappa \) is 0.1, 1, and 1 for AIC, BIC and HQ respectively. In fact, for the sample sizes in the range from \( n = 20 \) to \( n \), BIC and HQ always selected ARIMA\((0, 1, 1)\), which is the one suggested by Wei (1990). It is interesting to see that AIC actually chose ARIMA\((0, 1, 1)\) except for the last three times, where ARIMA\((1, 0, 1)\) was chosen. Apparently the last few observations are influential for AIC, but not so for BIC and HQ in terms of selection. This does indicate instability of model selection (AIC here).

Now suppose that we were asked to provide one-step ahead forecast of the unemployed number sequentially beginning at the 16th month till the 60th month (for an example). Note that due to smaller sample sizes at the present situation, it is possible that there is bigger instability in model selection. If one forecasts according to the selected ARIMA model, the ASEP is 1382, 1279 and 1311 for AIC, BIC and HQ respectively. About 90% of times, AIC, BIC and HQ chose AR(1), while ARIMA(0,0,1),
ARIMA(0,0,2) and ARIMA(2,0,2) were also chosen. Note that the most popular model AR(1) here is quite different from the ARIMA(0,1,1) model preferred by model selection when the sample sizes are between \(n - 20\) and \(n\). For AFTER (with weighting starting at the 10th month), the ASEP is 1135, a 9% reduction compared to the best of the model selection methods.

5.4 Data set 4

The data consist of Australian clay brick monthly production statistics (in million) aggregated into quarters from March 1956 to September 1994 (see, e.g., Makridakis, Wheelwright and Hyndman (1998, Chapter 1)).

Out of the \(n = 155\) observations, the last \(n_0 = 60\) were used to assess performance of the methods. ARIMA\((p, d, q)\) models with \(p, q = 0, ..., 5; d = 0, 1\) are considered as candidate models here. For AFTER, the weight begins at 85. The sequential instability \(\kappa\) (with \(L = 60\)) for AIC, BIC and HQ is 0.41, 0.24, and 0.18 respectively, suggesting that there is substantial selection instability in the data. The ASEP’s for AIC, BIC, HQ, and AFTER are 721.2, 826.5, 805.5, and 656.5 respectively. Note that AIC, BIC and HQ have 10%, 26% and 23% higher error compared to AFTER, respectively.

We also considered a log transformation. Though it has some improvements in diagnostic plots (residuals, ACF, goodness of fit, etc.), model selection is still instable with \(\kappa\) values for AIC, BIC and HQ being 0.18, 0.41, and 0.37 respectively. The perturbation plot in Section 2.4 also points at this direction. The ASEP’s of AIC, BIC and HQ are 8%, 22% and 18% higher compared to AFTER for the transformed data.

6 Concluding remarks

Time series models of the same type are often considered for fitting time series data. The task of choosing the most appropriate one for forecasting can be very difficult. In this work, we proposed the use of a combining method, AFTER, to convexly combine the candidate models instead of selection one of them. The hope is that when there is much uncertainty in finding the best models as is the case in many applications, combining may reduce instability of the forecast and therefore improve prediction accuracy. Simulation and real data examples do indicate potential advantage of AFTER over model selection for such cases.

Simple stability (or instability) measures were proposed for model selection. They can be used to give one an idea whether or not one can trust the selected model and the corresponding forecast when using a model selection procedure. If there is apparent instability, it is perhaps a good idea to consider combining the models.

Clearly, it is not a good idea to blindly combine all possible models. Preliminary analysis (e.g., examining ACF) should be performed to obtain a list of reasonable models. Transformation and differencing may also be considered in that process.
7 Acknowledgments

This research was supported by the United States National Science Foundation CAREER Award Grant DMS0094323.

References