REGRESSION-BASED FORECAST COMBINATION METHODS

Xiaoqiao WEI

Abstract

Least squares combinations (Granger & Ramanathan, 1984) are an important development in the forecast combination literature. However, ordinary least squares methods often perform poorly in real application due to the variability of coefficient/weight estimations. In this work, on one hand, we propose sequential subset selections to reduce the variability during combinations. On the other hand, we propose a novel method to simultaneously stabilize and shrink the coefficient/weights estimates. The proposed methods can be applied to various combination methods to improve prediction as long as their weights are determined based on ordinary least squares.

Keywords: forecast combinations, least squares, sequential selection, stabilization, shrinkage

JEL Classification: C32, E24

1. Introduction

In their original papers, Bates and Granger (1969), and Newbold and Granger (1974) showed that combined forecasts may reduce prediction variability under the conditions that the forecasts are unbiased. They proposed several combining rules based on estimated variance-covariances of the forecast candidates. Granger and Ramanathan (1984) later expanded these variance-covariance methods in a regression framework.

They argued that the variance-covariance methods can be treated as the least squares solutions under two constraints: one is that there is no constant term in the least squares formulation, the other is that all coefficients/weights are nonnegative and sum up to 1. They advocated use of regression methods that loosen these constraints for smaller mean squared prediction errors.

Since then, many regression-based forecast combination methods have been proposed in the literature. For example, Diebold (1988) considered serial correlation in the least squares framework. Coulson and Robins (1993) included a lagged dependent variable besides the forecast candidates. Deutsch *et al*. (1994) addressed regime switches when estimating coefficients/weights. Interested readers are referred to Timmermann (2006) and references therein.

Recently, researchers have worked on forecast combinations of a large number of forecasts in hope to take advantages of many different sources or models (e.g., Chan *et al*., 1999, Stock & Watson, 2003 and 2004, and Rapach & Strauss, 2005 and 2008). It has been shown, however, that the ordinary regression combination is not optimal for this kind of scenarios due to high variance. The empirical evidence of poor performances of the large-number regression combinations is provided by Rapach & Strauss (2005, 2008), among many others.

Chan *et al*. (1999) proposed the use of James-Stein estimation, ridge regression, and principle components regression as alternatives to ordinary least squares. Swanson and Zeng (2001) proposed regression combinations based on subset selections using AIC (Akaike, 1973) or BIC (Schwarz, 1978) to choose the best subset of all possible forecast candidates. Assume that there are p forecast candidates, by Swanson and Zeng's method, one has to select among in total $2p-1$ models, which is not practical when p is relatively large. Alternative to using AIC or BIC, one may apply subset selections by using t-statistic (e.g., Swanson and Zeng, 2001). However, this method often performs poorly in real applications (e.g., Rapach & Strauss, 2005 and 2008).

In this work, we first consider sequential subset selections in contrast to all subset selections, which reduce the number of models fitted to be at most $p(p+1)/2$. Simulations and real data examples show that sequential subset selections substantially improve upon ordinary least squares. Although sequential subset selections are commonly used to choose explanatory variables or orders in ARIMA modeling (see Zou & Yang, 2004, for related issues), to our knowledge, they have not been discussed in the forecast combination framework. From our numerical studies, sequential subset selections are a valuable technique for large-number forecast combinations.

We also propose a novel regression-based combination method, the decreasingly averaging method. Sequential subset selections discard insignificant forecast candidates using AIC or BIC. In contrast, the decreasingly averaging method retains all forecast candidates, but simultaneously stabilizes and slowly shrinks their coefficients/weights according to their order of appearance in the process of sequential selections. The less significant is the candidate, the more to be shrunk, thus the less effect on the combined forecasts. This is different from the existing Bayesian shrinkage methods (Stock & Watson, 2004), which shrink towards equal weights. Sequential subset selections and the decreasingly averaging method can be easily implemented. Actually, they can be tools to help other combination methods improve prediction accuracy especially in the large-number combination cases as long as their weights are determined based on ordinary least squares. For instance, it has been pointed out that Bayesian shrinkage methods have variable performance across different occasions or forecast horizons (Stock & Watson, 2004, and Rapach & Strauss, 2005 and 2008). In one real example which follows in section 4, we show that the proposed methods can help Bayesian shrinkage methods perform more stably and competitively compared to other combination methods.

Recently, Hansen (2008) proposed Mallows Model Averaging (MMA) for forecast combinations and found that the MMA method compared favorably with other feasible forecasting methods in terms of the one-step-ahead mean squared forecast error.

However, it is not applicable in the present setting of nonnested forecasting models, which are often encountered in real applications.

Yang (2004) pointed out there are two main directions of forecast combinations in the literature: combining for adaptation and combining for improvement. The first one targets the best individual performance among the pool of forecast candidates. The second one aims at significantly outperforming each individual forecast candidate. The early variance-covariance and regression based combination methods (including the proposed methods in this work) fall in the second direction. Interested readers are referred to Wei and Yang (2008) and references therein for some relevant work in the first direction. Different Bayesian combination methods can be categorized into either the first direction (e.g., Wright, 2003) or the second direction (e.g., Palm & Zellner, 1992).

The rest of the paper is organized as follows. In section 2, we propose sequential subset selections and the decreasingly averaging method. In section 3, we present simulation results for the proposed methods. In section 4, the proposed methods are examined through three data examples. Concluding remarks are given in section 5.

2. Methodologies

In this section, we first propose sequential subset selections and then the decreasingly averaging method. We shall also discuss the performance measures that will be used in next simulations and real data examples.

2.1. Sequential subset selections

Assume that there is a time series which we are interested in for forecasting, y_t, t = 1, 2, · · · , and there are p forecast candidates, x_{1t}, x_{2t}, · · · , x_{pt}, t = 1, 2, · · · . Granger and Ramanathan (1984) suggested the forecast combination:

$$y_{t+h} = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{it} + \varepsilon_{t+h}$$

where *h* is the forecast horizon, and the coefficients/weights can be estimated by least squares, possibly under some constraints. Then the combined forecast is given by

$$x_t^c = \hat{\alpha}_0 + \sum_{i=1}^{p} \hat{\alpha}_i x_{it} \quad ,$$

When p is large relative to the forecast sample size, the total variability of the coefficients/weights estimations is substantial and thus hurts the performance of the combined forecast x_t^c . A natural solution to this issue is to select the subset of the most significant forecast candidates, discarding others, in the regression formulation. The most commonly used selection criteria are AIC and BIC. AIC measures the discrepancy between the true model and a fitted model, while BIC approximates the posterior probabilities in a Bayesian framework. Assume that the number of parameters in the fitted model is k, and the sample size is n. AIC and BIC are both of the form

-log(maximized likelihood) + penalty,

where the penalty is k in AIC, or k*log(n)/2 in BIC. The model that minimizes the criterion is selected.

When p is large, direct applications of AIC or BIC over all subset models is infeasible. Sequential selections are a practical approach to proceed. To apply sequential subset selections to the p forecast candidates, we first choose the most significant one which minimizes AIC or BIC among the p candidates. Then we update the previous model by adding the second most significant one among the remaining p−1 candidates. If the AIC (or BIC) of the updated model is greater than that of the previous model, we stop and the previous model is our final choice. Otherwise, we continue to sequentially add one candidate a time until all of the p forecast candidates are exhausted.

2.2. The decreasingly averaging method

In this method, similarly to above, we first determine the most significant candidate using AIC or BIC, denoting it by $x_{(1)}$. Based on $x_{(1)}$, we determine the second most significant candidate, denoting it by $x_{(2)}$, and so on, until the least significant candidate, denoting it by $x_{(p)}$. Note that AIC and BIC provide exactly the same order of the p candidates. Then we fit the following p nested models:

$$y_{t+h} = \alpha_0 + \alpha_1 x_{(1)t} + \varepsilon_{t+h}$$

$$y_{t+h} = \alpha_0 + \alpha_1 x_{(1)t} + \alpha_2 x_{(2)t} + \varepsilon_{t+h}$$

$$\cdots$$

$$y_{t+h} = \alpha_0 + \alpha_1 x_{(1)t} + \alpha_2 x_{(2)t} + \alpha_3 x_{(3)t} + \ldots + \alpha_p x_{(p)t} + \varepsilon_{t+h}.$$

We obtain a p by (p + 1) matrix of the estimated coefficients:

$$\begin{pmatrix} \hat{\alpha}_0^1 & \hat{\alpha}_1^1 & 0 & \ldots & 0 \\ \hat{\alpha}_0^2 & \hat{\alpha}_1^2 & \hat{\alpha}_2^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \hat{\alpha}_0^p & \hat{\alpha}_1^p & \hat{\alpha}_2^p & \ldots & \hat{\alpha}_p^p \end{pmatrix}$$

where the superscript i, i = 1, 2, · · · , p, denotes the ith nested model, and the ith row has p−i zeros to represent the coefficients of the candidates which are not in the model. We then take arithmetic averages over each of the columns, and denote them by

$$\begin{pmatrix} \tilde{\alpha}_0 & \tilde{\alpha}_1 & \tilde{\alpha}_2 & \ldots & \tilde{\alpha}_p \end{pmatrix}$$

Finally, the combining weights of the p candidates (plus an intercept) are given by the averaged coefficients. The intercept and the coefficients of the most significant candidate are stabilized through averaging the p models. Other coefficients are not

only stabilized but forced to slightly shrink. The shrinkage occurs when the zeros are taken into the average calculations. The degrees of shrinkage of the coefficients of the candidates are in some sense proportional to the degrees of their insignificance.

For instance, $\tilde{\alpha}_p = \frac{1}{p}\hat{\alpha}_p$. The final weight of the least significant candidate is 1/p of its coefficient in the ordinary regression combination. Therefore, the insignificant candidates are of less importance, but still playing a role, in the forecast combinations. Note that the decreasingly averaging method is very different from simply averaging the forecast candidates.

2.3. Performance measures

In our simulations, the time series y_t, t = 1, 2, · · · , is generated by a mean function plus a random error. The conditional mean squared h-step-ahead forecasting error $E(y_{t+h} - x_t^c)^2$ on current time t is actually the sum of the squared conditional bias and conditional variance:

$$E(y_{t+h} - x_t^c)^2 = (m_{t+h|t} - x_t^c)^2 + v_{t+h|t},$$

where mt+h|t is the mean, and $v_{t+h|t}$ is the error variance, conditional on current time t. Since the conditional variance $v_{t+h|t}$ is always the same no matter which combination method is used for prediction, when comparing the performance of different combination methods, we may only consider the squared conditional bias as the net loss:

$$L(m_{t+h|t}, x_t^c) = (m_{t+h|t} - x_t^c)^2.$$

Accordingly, the corresponding net risk may be defined by $E(m_{t+h} - x_t^c)^2$. When evaluating the performance of a series of combined forecasts x_t^c, t = 1, 2, · · · , l, we consider the average forecasting risk

$$\text{Ave. Risk} = \frac{1}{1}\sum_{t=1}^{l} E(m_{t+h|t} - x_t^c)^2$$

This risk will be used as the performance measure in the following simulation investigations. In real data applications, since the above risk is unknowable, we instead consider the mean square prediction error

$$MSE = \frac{1}{l}\sum (y_{t+h} - x_t^c)^2$$

3. Simulations

In this section, we extensively examine the performances of the proposed methods under random model settings. We do not limit our focus on some individual models. In contrast, we evaluate the behavior of the proposed methods on a number of randomly

generated models. The simulation results under random model settings can provide us more fair and informative understanding of the proposed methods than those of some specific models.

We consider two kinds of scenarios. One is that the random model has a fixed order, and the other is that the random model has various orders. Under each scenario, we consider two cases. One is that the true model is in the forecast candidate set, and the other is that the true model is not. Sequential subset selections and the decreasingly averaging method can be easily applied to various combination methods based on ordinary least squares for prediction improvement. For instance, in Coulson and Robin's (1993) method, where the coefficients/weights of the forecast candidates plus a lagged dependent variable are determined by ordinary least squares, one can apply sequential subset selections to leave out insignificant terms to reduce the total estimation variability, or the decreasingly averaging method to simultaneously stabilize and shrink the coefficients/weights to improve upon the original versions. In this and following sections, we focus on applying the proposed methods to the ordinary regression combinations and Bayesian shrinkage methods.

Stock and Watson (2004) proposed Bayesian shrinkage methods as follows

$$w_i = \lambda \hat{\alpha}_i + (1-\lambda)(1/p)$$

where wi is the weight of the ith candidate, and α̂i is the estimated coefficient of the ith candidate in an ordinary regression combination which does not include an intercept. λ is a shrinkage tuning parameter, and is equal to max{0, 1−κ[p/(n−p)]}. In the following simulations and data examples, we consider κ equal to 0.5 and 1 respectively.

When k is large (λ is small), the weight shrinks toward equal weights from the least squares estimation. Diebold and Pauly (1990) pointed out that this kind of weight wi can be interpreted as a Bayesian estimator.

3.1. Random models with a fixed order

In this sub-section, we first consider that the true model has a fixed order AR(4) with 9 candidate models which are a white noise, AR(1) to AR(4), and MA(1) to MA(4), respectively. Note that the true model is in the candidate set. For this and the following simulations, the error term of the true model follows a normal distribution with mean zero and variance 4. We generate 100 models with coefficients randomly generated from the uniform distribution on [−1, 1] (non-stationary coefficients are discarded). We replicate each model and the candidate forecasts 100 times to simulate the forecasting risks. In each replication, we generate a sample with size 140. The candidate models start to generate one-step-ahead forecasts after 80 observations, then recursively do so once every additional observation is made. Thus there are in total 60 forecasts for each model. The combination methods use 40 forecasts to start weighting, evaluated on last 20 observations.

We consider two types of ordinary least squares combinations. One is with recursive fitted sizes, which start from 40 and sequentially increase by one until 59. The other is with rolling fitted sizes, which always use the most recent 40 forecasts. We denote the two methods by re-OLS and ro-OLS respectively. Furthermore, we consider two

Bayesian shrinkage methods as described above (denoted by Shk-1 and Shk-2 respectively). We apply the four methods to determine the combination weights of the 9 forecast candidates. We also apply the proposed methods to these four methods. When applying AIC or BIC selection, some insignificant forecast candidates are left out, and the four existing methods determine the coefficients/weights based on the remaining forecast candidates. When applying the decreasingly averaging method (denoted by DA), all of the 9 forecast candidates remain in the four existing methods, but their (OLS) coefficients are adjusted by DA. Table 1 shows the results of the comparisons. The second column gives the means of the 100 forecasting risks of the four existing combination methods. The rest of the columns give respectively the means of the 100 forecasting risks of the proposed methods applied to the existing methods. The numbers in parentheses are the improvement percentages of the proposed methods over the existing methods.

Table 1

Comparisons when the true model AR(4) is in the candidate set

		DA	AIC	BIC
re-OLS	1.53	0.82 (46%)	0.84 (45%)	0.69 (55%)
ro-OLS	2.12	1.01 (52%)	0.98 (54%)	0.79 (63%)
Shk-1	1.08	0.60 (45%)	0.61 (44%)	0.51 (53%)
Shk-2	1.02	0.66 (35%)	0.58 (43%)	0.50 (51%)

From Table 1, the recursive OLS outperforms the rolling OLS, while Bayesian shrinkage methods outperform the recursive OLS. All of the proposed methods can significantly improve upon the existing combination methods. When the true model is in the candidate set, BIC selection performs the best among the proposed methods, while the decreasingly averaging method performs similarly as AIC selection.

We then consider that the random true model has a fixed order AR(6) with the same 9 candidate models as in the previous experiment. Note that in this experiment the true model is not in the candidate set. Other settings remain the same as before. Table 2 shows the results of the comparisons.

The numbers in Table 2 are substantially bigger than those in Table 1 because in this experiment the true model is not in the candidate set. Every combination method produces bigger forecasting risks in this situation. The proposed methods, however, still significantly improve upon the existing methods. Furthermore, there are a couple of interesting phenomena worth mentioning.

Table 2

Comparisons when the true model AR(6) is not in the candidate set

		DA	AIC	BIC
re-OLS	3.18	2.28 (28%)	2.68 (16%)	2.63 (17%)
ro-OLS	3.80	2.44 (36%)	2.93 (23%)	2.77 (27%)
Shk-1	2.62	2.07 (21%)	2.40 (8%)	2.46 (6%)
Shk-2	2.53	2.19 (14%)	2.35 (7%)	2.43 (4%)

When the true model is not in the candidate set, the decreasingly averaging method performs the best among the proposed methods, while AIC and BIC selections perform similarly. In reality, it is usually the case where no model in the candidate set truly describes the underlying data generating process. Thus one may usually have many candidate models to entertain but without the true model. From this experiment, the decreasingly averaging method shows potential advantages in real applications.

3.2. Random models with various orders

In previous sub-section, the random true model has a fixed order (either AR(4) or AR(6)). In this sub-section, we consider the true model has different orders since in reality the underlying data generating process may have a structural change. We first consider that the true model uniformly varies from AR(1) to AR(4), while the candidate models are the same as in previous experiments. Note that even though the true model varies, it is still in the candidate set. Other settings remains the same as in the previous experiments. Table 3 shows the results of the comparisons.

It is of interest to compare Table 3 with Table 1. Even though the four existing methods yield different means of forecasting risks in the two tables, the proposed methods make similar improvements over the existing methods. Again in Table 3, when the true model is in the candidate set, BIC selection performs the best among the proposed methods, while the other two perform similarly.

Table 3

Comparisons when the true model varies and is in the candidate set

		DA	AIC	BIC
re-OLS	1.40	0.77 (45%)	0.76 (46%)	0.62 (56%)
ro-OLS	1.91	0.95 (50%)	0.90 (53%)	0.72 (62%)
Shk-1	0.95	0.54 (44%)	0.54 (44%)	0.44 (54%)
Shk-2	0.86	0.55 (36%)	0.52 (39%)	0.43 (50%)

We then consider that the random true model uniformly varies from AR(5) to AR(7) with the same 9 candidate models. Note that the true model is not in the candidate set. Other settings remain the same as before. Table 4 shows the results of the comparisons.

Table 4

Comparisons when the true model varies and is not in the candidate set

		DA	AIC	BIC
re-OLS	2.64	1.87 (29%)	2.17 (18%)	2.10 (20%)
ro-OLS	3.14	2.01 (36%)	2.34 (26%)	2.22 (29%)
Shk-1	2.15	1.68 (22%)	1.94 (10%)	1.93 (10%)
Shk-2	2.13	1.83 (14%)	1.90 (11%)	1.90 (11%)

From Table 4, the proposed methods significantly improve upon the four existing methods. As in Table 2, when the true model is not in the candidate set, AIC and BIC selections perform similarly, while the decreasingly averaging method significantly

outperforms both of them, which again shows its potential advantages in real applications.

So far we have dealt with the cases where the random error of the true model follows a normal distribution with mean zero and variance 4. Alternatively, we also consider the random error taking different variances from 0.5 to 7 and different distributions such as shifted gamma (with mean zero), double exponential, and t. We obtain similar results as presented in this paper, which are available upon request.

4. Data examples

In this section, we apply the proposed methods to three real data sets with a focus on the third one where we compare the proposed methods with other existing combination methods across different forecast horizons.

4.1. Data set 1

The data with n = 98 are levels of Lake Huron measured in each July from 1875 through 1972 (Brockwell & Davis, 1991). Graphical inspection suggests differencing the data. The candidates are ARMA(p,q) models with p, q = 0, 1, 2. The training sample size for the candidate models is 57. Then we obtain 40 one-step-ahead forecasts for each model. The combination methods use the beginning 20 forecasts to calculate the initial coefficients/weights. We compare the performance of the combination methods over the last 20 observations. Table 5 gives the comparison results. The second column gives the MSEs of the existing combination methods. The rest of the columns give respectively the MSEs of the proposed methods applied to the existing methods.

Table 5

Comparison results of data set 1

		DA	AIC	BIC
re-OLS	1.16	0.78 (33%)	0.76 (35%)	0.82 (30%)
ro-OLS	1.74	1.07 (38%)	1.04 (40%)	0.73 (58%)
Shk-1	0.85	0.73 (15%)	0.77 (9.8%)	0.85 (0.4%)
Shk-2	0.72	0.68 (4.9%)	0.76 (-5.9%)	0.84 (-17%)

In Table 5, the three proposed methods dramatically improve upon the recursive OLS method, and they perform comparably. The proposed methods also dramatically improve upon the rolling OLS method with BIC selection standing out. The decreasingly averaging method incorporates Bayesian shrinkage methods favorably compared to sequential subset selections.

4.2. Data set 2

The data are aggregated Australian clay brick quarter productions (in million units) from March 1956 through September 1994 (Makridakis *et al*., 1998). The data set consists of 155 observations. After taking a log transformation, we difference the data to improve the stationarity. The candidates are ARMA(p,q) models with p, q = 0, 1, 2,

3, 4, 5 (discard the case if the AR parts are not stationary). We obtain 20 candidate models. The training sample size for the candidates' models is 100. There are 54 one-step-ahead forecasts for each model. The combination methods use the beginning 34 forecasts to calculate the initial coefficients/weights. We compare the performance of the combination methods over the last 20 observations. Table 6 gives the comparison results (MSE×103).

Table 6

Comparison results of data set 2

		DA	AIC	BIC
re-OLS	6.5	4.9 (25%)	5.3 (18%)	6.7 (-3.1%)
ro-OLS	37.2	7.4 (80%)	5.4 (85%)	5.3 (86%)
Shk-1	5.4	5.1 (5.6%)	5.7 (-5.6%)	4.4 (19%)
Shk-2	6.0	6.0 (0.0%)	5.6 (6.7%)	4.4 (27%)

In Table 6, most of the proposed methods significantly outperform the two regression combination methods. For this data set, BIC selection incorporates Bayesian shrinkage methods favorably, while the decreasingly averaging method make an improvement by 5.6% or remains the same performance as the original Bayesian shrinkage method.

4.3. Data set 3

Rapach and Strauss (2005) studied the large-number combinations for forecasting employment growth in Missouri using 22 candidate models. They examined in total 20 different combination methods, where the recursive and rolling OLS and Bayesian shrinkage methods are all included (for more details about the data set, forecast candidates, and combination methods, please check their article). The Missouri employment growth data set spans from January 1976 to January 2005 and the combination methods are evaluated over the last 10 years. There are four forecast horizons considered, 3, 6, 12, and 24 months. Before we discuss the proposed methods, we present the MSEs of the best candidate, best combination, and simple average across the four horizons in Table 7.[1] To be conformable to Rapach and Strauss (2005), the entries in Table 7 and the following Table 8 are ratios of the MSEs of the methods to that of an AR benchmark model.

Table 7

The MSEs of some methods of data set 3

	Best ind.	Best com.	Simple average
h=3	0.90	0.94	0.96
h=6	0.83	0.91	0.92
h=12	0.70	0.71	0.84
h=24	0.76	0.51	0.83

[1] We rewrote the whole program in R statistical software, and found that there are very minor differences between our numerical results and those of Rapach and Strauss (2005).

From Table 7, we can find that the simple average method performed very well at short horizons (h = 3 or 6), and the best combined forecast (which is Shk-2) significantly outperformed the best individual forecast candidate at h = 24. Table 8 shows the results when we apply the proposed methods to the recursive and rolling OLS and Bayesian shrinkage methods.

Table 8

Comparison results of data set 3 across different forecast horizons

			DA	AIC	BIC
	re-OLS	1.44	1.15 (20%)	1.24 (14%)	0.93 (36%)
h=3	ro-OLS	2.28	1.75 (23%)	2.05 (10%)	1.51 (34%)
	Shk-1	1.18	1.04 (12%)	1.18 (0.0%)	0.89 (24%)
	Shk-2	1.08	0.99 (9.1%)	1.13 (-4.6%)	0.89 (18%)
	re-OLS	1.56	1.24 (21%)	1.27 (19%)	1.10 (29%)
h=6	ro-OLS	2.45	1.89 (23%)	2.15 (12%)	1.68 (31%)
	Shk-1	1.21	1.01 (17%)	1.05 (13%)	0.97 (20%)
	Shk-2	1.02	0.91 (11%)	1.00 (2.0%)	0.96 (5.9%)
	re-OLS	1.27	1.11 (13%)	1.24 (2.4%)	1.13 (11%)
h=12	ro-OLS	2.81	2.13 (24%)	2.49 (11%)	2.30 (18%)
	Shk-1	0.85	0.79 (7.1%)	0.95 (-12%)	0.92 (-8.2%)
	Shk-2	0.71	0.71 (0.0%)	0.91 (-28%)	0.91 (-28%)
	re-OLS	0.95	0.75 (21%)	0.83 (13%)	0.75 (21%)
h=24	ro-OLS	1.62	1.28 (21%)	1.54 (4.9%)	1.25 (23%)
	Shk-1	0.51	0.55 (-7.8%)	0.54 (-5.9%)	0.52 (-2.0%)
	Shk-2	0.53	0.57 (-7.5%)	0.53 (0.0%)	0.52 (1.9%)

From Table 8, we can find that the proposed method can significantly improve upon the recursive and rolling OLS methods, and at h = 24, they perform very well, reaching 0.75. More interesting things happen to Bayesian shrinkage methods. We can find that Bayesian shrinkage methods have variable performance across different horizons. They performed very well at long horizons (h = 12 or 24), but poorly at short horizons (h = 3 or 6). However, BIC selection plus Bayesian shrinkage can reach 0.89 at h = 3, and the decreasingly averaging method plus Bayesian shrinkage can reach 0.91 at h = 6. The decreasingly averaging method plus Bayesian shrinkage makes improvement by 6.7% or remains the same as the original Bayesian shrinkage method at h = 12. The proposed methods plus Bayesian shrinkage methods have slightly worse performance than the original Bayesian shrinkage methods at h = 24, but still significantly outperform other combination methods. If we simply incorporate the decreasingly averaging method with the second Bayesian shrinkage method, we will reach 0.99, 0.91, 0.71, and 0.57, at the four horizons respectively, which makes the Bayesian shrinkage method the most attractive method out of the 20 combination methods across different forecast horizons.

5. Concluding remarks

The Least squares combinations (Granger & Ramanathan, 1984) are an important development in the forecast combination literature. The methods include the early variance-covariance methods as their special cases in some sense. Recently, researchers have worked on large-number forecast combinations. It has been shown that ordinary least squares combinations of all forecast candidates may have very poor performance in such situations. Due to computational difficulty, all subset selections are unattractive. As a solution, we propose two approaches, sequential subset selections and the decreasingly averaging method. The proposed methods are easily implemented, and can be tools to help various combination methods to improve prediction accuracy as long as their coefficients/weights are determined based on ordinary least squares. In this work, we focus on applying the proposed methods on the ordinary regression combinations and Bayesian shrinkage methods.

Sequential subset selections discard insignificant forecast candidates to reduce the variability of coefficient/weight estimations, leading to possibly improved predictions. The decreasingly averaging method retains all the candidates, but simultaneously stabilizes and slowly shrinks the coefficients/weights according to their significance, which is different from Bayesian shrinkage methods, which shrink towards equal weights.

We conduct structured simulations to examine the performance of sequential subset selections and the decreasingly averaging method. The numerical results show the proposed methods can significantly improve upon the recursive and rolling OLS and Bayesian shrinkage methods. When the true model is in the candidate set, BIC performs the best among the proposed methods, while the other two perform similarly. When the true model is not in the candidate set, the decreasingly averaging method significantly outperforms AIC and BIC selections, while AIC and BIC selections perform similarly. Three real data examples also confirm the potential advantages of the proposed methods. Especially, in data set 3, we examine their performance in a comprehensive setting, comparing them with 20 different combination methods. We find that the proposed methods can help Bayesian shrinkage methods improve prediction accuracy at short horizons. In particular, the decreasingly averaging method can help the second Bayesian shrinkage method be the most attractive combination method out of the 20 combination methods across different forecast horizons.

The theoretical understanding of the decreasingly averaging method remains future investigations. Another direction of future work is to examine the proposed methods on other combination methods based on ordinary least squares.

References

Akaike, H. (1973). “Information Theory and an Extension of the Maximum Likelihood Principle”. In: B. N. Petrov, & F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory*, Budapest: Akademia Kiado.

Bates, J. M. and Granger, C. W. J. (1969). "The combination of forecasts". *Operations Research Quarterly*, 20: 451-468.

Chan, Y. Z., Stock, J. H., and Watson, M. W. (1999). "A dynamic factor model framework for forecast combination". *Spanish Economic Review*, 1: 91-121.

Coulson, N. E. and Robins, R. P. (1993). "Forecasting combination in a dynamic setting". *Journal of Forecasting*, 12: 63-67.

Diebold, F. X. (1988). "Serial Correlation and the Combination of Forecasts". *Journal of Business & Economic Statistics*, 6: 105-111.

Deutsch, M., Granger, C. W. J., and Terasvirta,¨ T. (1994). "The Combination of Forecasts Using Changing Weights". *International Journal of Forecasting*, 10: 47-57.

Diebold, F. X. and Pauly, P. (1990). "The Use of Prior Information in Forecast Combination". *International Journal of Forecasting*, 6: 503-508.

Granger, C. W. J. and Ramanathan, R. (1984). "Improved methods of Forecasting". *Journal of Forecasting*, 3: 197-204.

Greene, W. H. (2000). *Econometric Analysis* (4th edition). New York: Prentice Hall.

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

Makridakis, S., Wheelwright, S., and Hyndman, R. J. (1998). *Forecasting: Methods and Applications* (3rd edition). New York: Wiley.

Newbold, P. and Granger, C. W. J. (1974). "Experience with Forecasting Univariate Time Series and the Combination of Forecasts". *JRSSA*, 137: 131-165.

Palm, F. C. and Zellner, A. (1992). "To Combine or not to Combine? Issues of Combining Forecasts". *Journal of Forecasting*, 11: 687-701.

Rapach, D. E. and Strauss, J. K. (2005). "Forecasting Employment Growth in Missouri with Many Potentially Relevant Predictors: An Analysis of Forecast Combining Methods". Federal Reserve Bank of St. Louis Regional Economic Development, 1: 97-112.

Rapach, D. E. and Strauss, J. K. (2008). "Forecasting US Employment Growth Using Forecasting Combining Methods". *Journal of Forecasting*, 27: 75-93.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics*, 6: 461-464.

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

Stock, J. H. and Watson, M. W. (2004). "Combination Forecasts of Output Growth in a Seven-Country Data Set". *Journal of Forecasting*, 23: 405-430.

Swanson, N. R. and Zeng, T. (2001). "Choosing among Competing Econometric Forecasts: Regression-based Forecast Combination Using Model Selection". *Journal of Forecasting*, 20: 425-440.

Timmermann, A. (2006). "Forecast Combinations". In: G. Elliott *et al*. (Eds.), *Handbook of Economic Forecasting*, Amsterdam: Elsevier.

Wei, X. and Yang, Y. (2008). "Robust Forecast Combinations". Submitted.

Wright, J. H. (2003). "Forecasting U.S. Inflation by Bayesian Model Averaging". International Finance Discussion Papers, No.780, Board of Governors of the Federal Reserve System.

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

Zou, H. and Yang, Y. (2004). "Combining Time Series Models for Forecasting". *International Journal of Forecasting*, 20: 69-84.