Journal of Econometrics 178 (2014) 294–309

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Is there an optimal forecast combination?☆

Cheng Hsiao [a,b,*], Shui Ki Wan [c]

[a] University of Southern California, United States
[b] WISE, Xiamen University, China
[c] Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Available online 14 November 2013

ABSTRACT

We consider several geometric approaches for combining forecasts in large samples—a simple eigenvector approach, a mean corrected eigenvector and trimmed eigenvector approach. We give conditions where geometric approach yields identical result as the regression approach. We also consider a mean and scale corrected simple average of all predictive models for finite sample and give conditions where simple average is an optimal combination. Monte Carlos are conducted to compare the finite sample performance of these and some popular forecast combination and information combination methods and to shed light on the issues of "forecast combination" vs "information combination". We also try to shed light on whether there exists an optimal forecast combination method by comparing various forecast combination methods to predict US real output growth rate and excess equity premium.

1. Introduction

A good forecasting model is of critical importance to investors for forming their portfolio decisions and to governments or business leaders for making policy decisions. Often there are a number of predictors for a variable of interest. Instead of focusing on the selection of the best forecasting model, Bates and Granger (1969) have suggested to combine different forecasts. The main arguments favoring combining forecasts are: (i) The true data generating process is unknown. Even the most complicated model is likely to be misspecified and can, at best, provide a reasonable "local" approximation. It is highly unlikely that a single model will dominate uniformly over time. (ii) The best model may change over time in a way that can be difficult to track on the basis of past forecasting performance. Combining forecasts across different models may be viewed as a way to make the forecast more robust against misspecification biases and measurement errors in the data set. (iii) It is possible that diversification gains from combining across a set of forecasting models will dominate the strategy of only using a single forecasting models. Since then, numerous forecast combination methods have been proposed (e.g. see the survey by Timmermann (2006)). In this paper we suggest some additional forecast combination methods. We evaluate the performance of these approaches with some popular forecast combination methods through Monte Carlo studies and two empirical applications of predicting real US GDP growth rate and excess equity premium on S&P 500.

Suppose y_t is a variable of interest and there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$, for y_t, for $t = 1, \ldots, T_1$. Our goal is to find a linear combination $\mathbf{w} = (w_1, \ldots, w_N)'$, an $N \times 1$ vector of constants to form a new predictor $\widehat{y}_t = \mathbf{w}'\mathbf{f}_t, t = T_1+1, \ldots, T$ which is optimal in terms of the risk function chosen. Obviously, the choice of $\mathbf{w}$ depends on the loss function of the prediction error and the available sample.

Our objective is to find $\mathbf{w}$ to minimize the mean squared prediction error (*MSPE*)

$$\min E\,(y_t - \widehat{y}_t)^2\,. \tag{1}$$

The sample version of *MSPE* will be

$$MSPE = \frac{1}{T - T_1}\sum_{t=T_1+1}^{T}(y_t - \widehat{y}_t)^2\,. \tag{2}$$

In Section 2, we review some popular forecasting combination and information combination methods. In Section 3, we suggest an eigenvector approach to combine forecasts when sample size is large. Section 4 suggests a mean corrected and a trimmed eigenvector approach. Section 5 suggests a mean and scale corrected simple average method and gives conditions that simple average can be optimal. Section 6 compares the finite sample performance of the forecast combination methods suggested here vis-a-vie some

☆ The authors would like to thank Z.D. Bai, G. Elliott, C. Lu, M.H. Pesaran, A. Zellner, two referees and the Editor for helpful discussions and comments. Cheng Hsiao's work is partially supported by the National Science Foundation of China grant #71131008. Shui Ki Wan's work is partially supported by the RGC Grant # HKBU Project Code 255212.

* Corresponding author at: University of Southern California, United States.
E-mail addresses: chsiao@usc.edu (C. Hsiao), shuikiwan@gmail.com (S.K. Wan).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.11.003

popular sampling and Bayesian approach for combining forecasts. Section 7 compares the performance of various forecast combination methods to predict the US real output growth rate and the US excess equity premium. Concluding remarks are in Section 8.

2. Some popular forecast combination and information combination methods

2.1. Regression approach

Granger and Ramanathan (1984) suggest finding $\mathbf{w}$ to minimize (2) from a regression approach. They consider three regression models

$$y_t = \mathbf{w}'\mathbf{f}_t + u_t, \text{ subject to } \sum_{j=1}^{N} w_j = 1, \tag{3}$$

$$y_t = \mathbf{w}'\mathbf{f}_t + u_t, \tag{4}$$

$$y_t = \alpha + \mathbf{w}'\mathbf{f}_t + u_t. \tag{5}$$

which are named as *GR*1, *GR*2 and *GR*3, respectively.

*GR*3 yields an unconstrained regression weights and unconstrained minimum of $\sum_{t=1}^{T_1}(y_t - \widehat{y}_t)^2$. *GR*1 and *GR*2 can be viewed as constrained regression model of *GR*3 with *GR*1 being the most restrictive. If some of the predictive models f_{it} are biased predictors, then the weighting schemes generated by *GR*1 and *GR*2 could be biased, but *GR*3 still generates unbiased predictor as the bias in f_{it} is picked up by the intercept. Therefore, if sample size is large, we should expect the mean squared prediction error of *GR*3 $\leq$ mean squared prediction error of *GR*2 $\leq$ mean squared prediction error of *GR*1 unless the restriction is correct.

In addition to using regression method to find the weights, simple averages of all predictive models have been found to yield good forecasts (e.g. Clemen (1989) and Chan et al. (1999)). That is, let

$$w_i^{SA(N)} = \frac{1}{N}, \quad i = 1, \ldots, N. \tag{6}$$

We shall call forecasts generated by simple averaging of all predictive models *SA*(*N*).

Alternatively, Bates and Granger (1969) suggest finding the weight for the *i*-th predictive model as

$$w_i^{BG} = \frac{\widehat{\sigma}^{-2}(i)}{\sum_{j=1}^{N}\widehat{\sigma}^{-2}(j)}, \quad i = 1, \ldots, N, \tag{7}$$

where $\widehat{\sigma}^2(i) = \frac{1}{T_1}\sum_{t=1}^{T_1}(y_t - f_{it})^2$ is the estimated mean squared prediction error of the *i*th model.

Both (6) and (7) may be viewed as the constrained solution to the regression model (3) with prior knowledge that $w_i = w_i^{SA(N)} = \frac{1}{N}$ or $w_i = w_i^{BG}$, for $i = 1, \ldots, N$.

2.2. Bayesian averaging

Buckland et al. (1997) propose to find the weights through a Bayesian procedure. Given the prior that any of the *N* predictive models is equally likely to be the best model for forecasting y_t, they show that[1]

$$w_{BIC,i} = \frac{\exp\left(-\frac{1}{2}\Delta BIC_i\right)}{\sum_{j=1}^{N}\exp\left(-\frac{1}{2}\Delta BIC_j\right)}, \quad i = 1, \ldots, N, \tag{8}$$

gives the posterior odds for the *i*th model being the best predictive model where $\Delta BIC_i = BIC_i - \min_j\left(BIC_j\right)$, and BIC_i is the Bayesian information criterion for the *i*th model (Schwartz, 1978),

$$BIC_i = T_1 \ln\widehat{\sigma}^2(i) + (m_i + 1)\ln(T_1), \tag{9}$$

where m_i denotes the number of unknown parameters in the *i*th predictive model. We shall denote the Bayesian averaging methods as *MABICFC*.

2.3. Model selection approach

Instead of trying to find the optimal weight for each of the forecasting model. Swanson and Zeng (2001) propose to treat the predicted value of each forecasting model as a regressor, then use model selection criterion to select the optimal combination out of $2^N - 1$ possible combinations. Therefore, if the predicted values of *m* predictive models are used as regressor, then $m_i = m$. We shall consider the choice in terms of *BIC*, *AIC* (Akaike, 1974), and *AICC* (Hurvich and Tsai, 1989)

$$AIC_i = T_1 \ln\widehat{\sigma}^2(i) + 2\left(m_j + 1\right), \tag{10}$$

$$AICC_i = AIC_i + 2\frac{(m_i + 1)(m_i + 2)}{T_1 - m_i - 2}. \tag{11}$$

2.4. Information combination

Suppose the information set for generating the predictive models, $\mathbf{x}$, is known. Let the mean squared error of $\left(y_t, \mathbf{x}_t'\right)'$ be denoted as

$$\begin{pmatrix}\sigma_y^2 & \sigma_{yx}\\ \sigma_{xy} & \Sigma_{xx}\end{pmatrix}. \tag{12}$$

We consider three information combination approaches. The first is the regression approach. Then the minimum mean squared error predictor for y_t is

$$\widehat{y}_t = \mathbf{x}_t'\Sigma_{xx}^{-1}\sigma_{xy}, \tag{13}$$

which will be denoted as *TVC*.

When Σ_{xx} and σ_{xy} are unknown, we replace them by the corresponding sample estimates and denote the resulting predictor as *EVC*.

The second approach is to use one of the model selection criteria to select the best subset of $\mathbf{x}$ to generate predictions. Again, we use *AIC*, *AICC* and *BIC* criteria.

The third is to use Bayesian averaging method knowing how the predictive models are generated. We shall call it *MABIC*, where the averaging weights given by (8) are in terms of *BIC*.

3. An eigenvector approach

Let $\mathbf{v}_t$ be the $N \times 1$ prediction error vector of $\mathbf{f}_t$ with the *i*th element being the prediction error of the *i*th predictive model,

$$v_{it} = y_t - f_{it}, \quad i = 1, \ldots, N. \tag{14}$$

Our goal is to find a linear combination of v_{it} such that the mean squared prediction error

$$E\left(\mathbf{w}'\mathbf{v}_t\right)^2 = \mathbf{w}'E\left(\mathbf{v}_t\mathbf{v}_t'\right)\mathbf{w} = \mathbf{w}'\Sigma\mathbf{w} \tag{15}$$

is minimized subject to some normalization condition, where Σ is the mean squared prediction error matrix of $\mathbf{f}_t$. The conventional normalization condition is (e.g. Markowitz (1952, 1959), Newbold and Granger (1974) and Timmermann (2006)),

$$\mathbf{e}'\mathbf{w} = \sum_{i=1}^{N} w_i = 1, \tag{16}$$

[1] Diebold and Pauly (1990) considered the use of prior information in forecast combination.

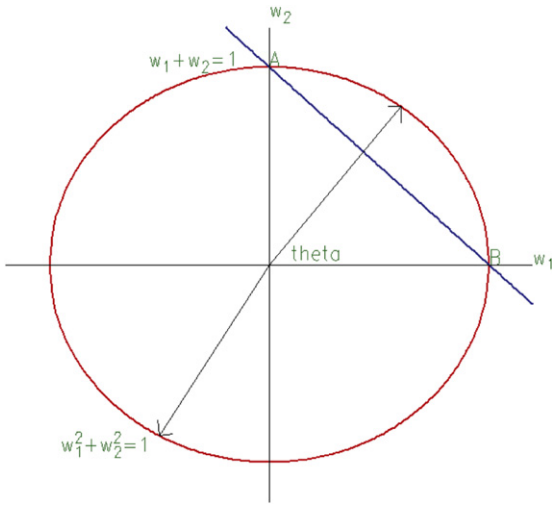

Fig. 1. Eigenvector approach.

and $w_i \geq 0, i = 1, \ldots, N$, where $\mathbf{e}$ is an $N \times 1$ vector of $(1, \ldots, 1)'$. Minimizing (15) subject to (16) yields

$$\mathbf{w}^{VC} = \left(\mathbf{e}'\Sigma^{-1}\mathbf{e}\right)^{-1}\left(\Sigma^{-1}\mathbf{e}\right). \tag{17}$$

However, the normalization condition for (15) is not an innocuous condition. The normalization condition (16) puts a restriction on w_i, which leads to a constrained minimization of (15). Therefore, instead of (16), we propose to use the normalization condition,

$$\mathbf{w}'\mathbf{w} = \sum_{i=1}^{N} w_i^2 = 1. \tag{18}$$

Minimizing (15) subject to $\mathbf{w}'\mathbf{w} = 1$ leads to $\mathbf{w} = \mathbf{w}^1$ where $\mathbf{w}^1$ is the eigenvector corresponding to the smallest eigenvalue of Σ, say ϕ_1.

Minimizing (15) subject to $\mathbf{e}'\mathbf{w} = 1$ restricts the search on a R^{N-1} hyperplane in R^N, while the normalization $\mathbf{w}'\mathbf{w}= \mathbf{1}$ allows the search of the minimum of (15) in R^N. This can be seen by considering a simple case of $N = 2$ (see Fig. 1). The relative weight of w_1 and w_2 is given by $\tan(\theta)$ where θ is the angle between the straight-line going through the origin and the w_1 axis. The constraint $w_1 + w_2 = 1$ and $w_i \geq 0$ limit the search of (w_1, w_2) along the straight line AB (the first quadrant) on the Figure. Hence, the relative weight of w_2 and w_1 is given by the angle between the straight line going through the origin and the point on AB and w_1 axis. On the other hand, the normalization rule $w_1^2 + w_2^2 = 1$ does not impose any constraint. It allows the search of the minimum over the complete plane (360 degree). In other words, any $\mathbf{w}^*$ that satisfies $\mathbf{e}'\mathbf{w}^* = 1$ can be converted to the constraint $\widetilde{\mathbf{w}}^{*\prime}\widetilde{\mathbf{w}}^* = 1$ by letting $\widetilde{\mathbf{w}}^* = \frac{1}{\sqrt{c}}\mathbf{w}^*$, where $c = \mathbf{w}^{*\prime}\mathbf{w}^*$. However, not every $\widetilde{\mathbf{w}}^{**}$ that satisfies $\widetilde{\mathbf{w}}^{**\prime}\widetilde{\mathbf{w}}^{**} = 1$ can be converted to $\mathbf{e}'\mathbf{w}^{**} = 1$ where $\mathbf{w}^{**}$ is a constant multiple of $\widetilde{\mathbf{w}}^{**}$. Therefore, in principle, minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{w}'\mathbf{w} = 1$ should yield a smaller value than subject to $\mathbf{e}'\mathbf{w} = 1$ when both are compared on the same scale.

We illustrate this point by considering the following two examples:

Example 3.1. Suppose

$$\Sigma = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}.$$

Minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ or $\mathbf{w}'\mathbf{w} = 1$ yields $\mathbf{w}^{*\prime} = \left(\frac{1}{2}, \frac{1}{2}\right)$ or $\widetilde{\mathbf{w}}^{**\prime} = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ or $(\widetilde{\mathbf{w}}^{**\prime} = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right))$, respectively. Let $\widetilde{\mathbf{w}}^{*\prime} = \sqrt{2}\mathbf{w}^*$, we have $\widetilde{\mathbf{w}}^{*\prime}\widetilde{\mathbf{w}}^* = 1$. However, $\widetilde{\mathbf{w}}^{*\prime}\Sigma\widetilde{\mathbf{w}}^* = 0.5 > \widetilde{\mathbf{w}}^{**\prime}\Sigma\widetilde{\mathbf{w}}^{**} = 0.4$.

Example 3.2. Suppose

$$\Sigma = \begin{pmatrix} 1.4 & -0.4 \\ -0.4 & 1.4 \end{pmatrix}.$$

Minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ or $\mathbf{w}'\mathbf{w} = 1$ yields $\mathbf{w}^{*\prime} = \left(\frac{1}{2}, \frac{1}{2}\right)$ or $\widetilde{\mathbf{w}}^{**\prime} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$. Let $\mathbf{w}^{**\prime} = \frac{1}{\sqrt{2}}\mathbf{w}^{**\prime}$, then $\mathbf{e}'\mathbf{w}^{**} = 1$, and $\mathbf{w}^{*\prime}\Sigma\mathbf{w}^* = \mathbf{w}^{**\prime}\Sigma\mathbf{w}^{**} = 0.5$.

Example 3.1 shows that it is possible to rescale $\mathbf{w}^*$ into $\widetilde{\mathbf{w}}^*$ such that $\widetilde{\mathbf{w}}^{*\prime}\widetilde{\mathbf{w}}^* = 1$, but it is not possible to rescale $\widetilde{\mathbf{w}}^{**}$ into $\mathbf{w}^{**}$. In other words, when it is possible to rescale $\widetilde{\mathbf{w}}^{**}$ into $\mathbf{w}^{**}$ such that $\mathbf{e}'\mathbf{w}^{**} = 1$, minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ or $\mathbf{w}'\mathbf{w} = 1$ yields identical minimum when the solution is normalized on the same scale. On the other hand, if it is not possible to renormalize the optimal solution subject to $\mathbf{w}'\mathbf{w} = 1$ into $\mathbf{e}'\mathbf{w}^{**} = 1$, then minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{w}'\mathbf{w} = 1$ will yield a solution no larger than the solution subject to $\mathbf{e}'\mathbf{w} = 1$.

However, minimizing (15) is not equivalent to minimizing the population version of *MSPE* (1). This can be seen by noting that

$$\begin{aligned} \mathbf{w}'\Sigma\mathbf{w} &= \mathbf{w}'E\left[\left(\mathbf{e}y' - \mathbf{F}'\right)\left(\mathbf{y}\mathbf{e}' - \mathbf{F}\right)\right]\mathbf{w} \\ &= E\left[\left(y_t^* - \mathbf{f}_t'\mathbf{w}\right)^2\right] \\ &= d^2E\left(y_t - \frac{1}{d}\mathbf{f}_t'\mathbf{w}\right)^2, \quad \text{when } d \neq 0, \end{aligned} \tag{19}$$

where $\mathbf{y}$ is a $T_1 \times 1$ vector of y_t, $\mathbf{F}$ is a $T_1 \times N$ matrix with the t-th row equal to $\mathbf{f}_t'$, $d = \mathbf{e}'\mathbf{w}$, and $y_t^* = dy_t$. Therefore, minimizing (1) is equivalent to minimizing $\frac{1}{d^2}\mathbf{w}'\Sigma\mathbf{w}$ over $\mathbf{w}$.

Since Σ is a positive definite matrix, we can arrange the N positive eigenvalues in increasing order $(\phi_1 = \phi_{\min}, \phi_2, \ldots, \phi_N = \phi_{\max})$. Let $\mathbf{w}^j$ be the eigenvector corresponding to ϕ_j. We propose to choose $\mathbf{w}$ corresponding to the minimum of $\left\{\frac{\phi_1}{d_1^2}, \frac{\phi_2}{d_2^2}, \ldots, \frac{\phi_N}{d_N^2}\right\}$, say $\mathbf{w}^l$, where $d_j = \mathbf{e}'\mathbf{w}^j$, then we set the combination weight as

$$\mathbf{w}^{EIG1} = \frac{1}{d_l}\mathbf{w}^l. \tag{20}$$

We shall refer this method of obtaining the weight from sample estimated mean squared prediction error matrix as *EIG*1.

Remark 3.1. Since Σ is usually unknown, if we substitute Σ by $\mathbf{S} = \frac{1}{T_1}\sum_{t=1}^{T_1}\left(\mathbf{e}y_t - \mathbf{f}_t\right)\left(y_t\mathbf{e}' - \mathbf{f}_t'\right)$ in (17), then the sample counterpart of (17) yields $\mathbf{w}^{VC} = \left(\mathbf{e}'\mathbf{S}^{-1}\mathbf{e}\right)^{-1}\mathbf{S}^{-1}\mathbf{e}$. This method will be named *VC*. The Bates and Granger (1969) method is equivalent to imposing a prior restriction that $\mathbf{S}$ is diagonal in the *VC* solution, $\mathbf{w}^{VC} = \left(\mathbf{e}'\mathbf{S}^{-1}\mathbf{e}\right)^{-1}\mathbf{S}^{-1}\mathbf{e}$.

Remark 3.2. Minimizing $\mathbf{w}'\mathbf{S}\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = \mathbf{1}$ transforms the norm back to minimization of the squared distance between y_t and $\widehat{y}_t = \mathbf{w}'\mathbf{f}_t$ along the y-axis. To see this, we note that

$$\begin{aligned} \mathbf{w}'\mathbf{S}\mathbf{w} &= \mathbf{w}'\left[\frac{1}{T_1}\left(\mathbf{e}y' - \mathbf{F}'\right)\left(\mathbf{y}\mathbf{e}' - \mathbf{F}\right)\right]\mathbf{w} \\ &= \frac{1}{T_1}\sum_{t=1}^{T_1}\left(y_t - \mathbf{w}'\mathbf{f}_t\right)^2, \end{aligned} \tag{21}$$

where $\mathbf{y}$ is a $T_1 \times 1$ vector of y_t, $\mathbf{F}$ is a $T_1 \times N$ matrix with the t-th row equal to $\mathbf{f}_t'$. In other words, minimizing (21) subject to $\mathbf{e}'\mathbf{w} = \mathbf{1}$ yields an identical result as *GR*1.

Remark 3.3. The difference between Granger and Ramanathan's (1984) regression approach of finding $\mathbf{w}$ and the eigenvector approach is that the former treats the predictive models, $\mathbf{f}_t$, as fixed and puts all the uncertainties on y_t. It finds $\mathbf{w}$ to minimize the

norm along the y-axis, hence it is sensitive to the outlying observations of y_t. On the other hand, the eigenvector approach is to find $\mathbf{w}$ from the so-called orthogonality principle (e.g. Golub and Van Loan (1980)). It treats uncertainties in y_t and the uncertainties of predictive models, $\mathbf{f}_t$, symmetrically. It aims at fitting a geometrically "best" subspace to the points $\mathbf{v}_t = \mathbf{e}y_t - \mathbf{f}_t$, $t = 1, \ldots, T_1$. Therefore, the eigenvector approach will be sensitive to the disparities of the performance of different predictive models.

4. A mean corrected and a trimmed eigenvector approach

The prediction error, v_{it}, may be decomposed as the sum of three components,

$$v_{it} = \alpha_i + \mathbf{b}_i'\boldsymbol{\lambda}_t + \varepsilon_{it}, \tag{22}$$

where α_i denotes the bias, $\boldsymbol{\lambda}_t$ denotes the $r \times 1$ omitted common factors of all predictive models, ε_{it} is the idiosyncratic component that is uncorrelated across i, $\mathbf{b}_i'$ denotes the $1 \times r$ constant factor loading vector that captures the impact of $\boldsymbol{\lambda}_t$ on the prediction error, v_{it}. Then the mean squared prediction error matrix can be expressed as

$$\Sigma = E\left(\mathbf{v}_t\mathbf{v}_t'\right) = \mathbf{B}\mathbf{B}' + \mathbf{D} + \boldsymbol{\alpha}\boldsymbol{\alpha}' \tag{23}$$

under the assumption that $E\left(\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t'\right) = \mathbf{I}_r$, where $\mathbf{B}$ is the $N \times r$ factor loading matrix with the ith row equals to $\mathbf{b}_i'$, $\mathbf{D}$ is the diagonal matrix with the ith diagonal element equals to $var\,(\varepsilon_{it}) = \sigma_i^2$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ is the $N \times 1$ vector.

If some of the predictive models yield biased predictions, $\boldsymbol{\alpha} \neq \mathbf{0}$. More accurate predictions can be obtained by eliminating the bias. We note that

$$\begin{aligned}
\mathbf{w}'\Sigma\mathbf{w} &= \mathbf{w}'E\frac{1}{T_1}\left\{\begin{array}{l}\left[\left(\mathbf{e}\mathbf{y}' - \mathbf{e}E\mathbf{y}'\right) - \left(\mathbf{F}' - E\mathbf{F}'\right) + \left(\mathbf{e}E\mathbf{y}' - E\mathbf{F}'\right)\right] \\ \left[\left(\mathbf{y}\mathbf{e}' - E\mathbf{y}\mathbf{e}'\right) - (\mathbf{F} - E\mathbf{F}) + \left(E\mathbf{y}\mathbf{e}' - E\mathbf{F}\right)\right]\end{array}\right\}\mathbf{w} \\
&= \mathbf{w}'E\left\{\frac{1}{T_1}\left[\left(\mathbf{e}\mathbf{y}' - \mathbf{e}E\mathbf{y}'\right) - \left(\mathbf{F}' - E\mathbf{F}'\right)\right]\right. \\
&\quad \left.\times\left[\left(\mathbf{y}\mathbf{e}' - E\mathbf{y}\mathbf{e}'\right) - (\mathbf{F} - E\mathbf{F})\right]\right\}\mathbf{w} \\
&\quad + \mathbf{w}'E\left\{\frac{1}{T_1}\left(\mathbf{e}E\mathbf{y}' - E\mathbf{F}'\right)\left(E\mathbf{y}\mathbf{e}' - E\mathbf{F}\right)\right\}\mathbf{w}.
\end{aligned} \tag{24}$$

The second term on the right-hand side of (24) is positive if $E\mathbf{F} \neq E\mathbf{y}\mathbf{e}'$. The first term on the right-hand side of (24) is equal to

$$\mathbf{w}'\Omega\mathbf{w} = \widetilde{d}^2E\left[(y_t - Ey_t) - \left(\mathbf{f}_t - E\mathbf{f}_t\right)'\frac{1}{\widetilde{d}}\mathbf{w}\right]^2, \tag{25}$$

where $\Omega = E\left\{\frac{1}{T_1}\left[\left(\mathbf{e}\mathbf{y}' - \mathbf{e}E\mathbf{y}'\right) - \left(\mathbf{F}' - E\mathbf{F}'\right)\right]\left[\left(\mathbf{y}\mathbf{e}' - E\mathbf{y}\mathbf{e}'\right) - (\mathbf{F} - E\mathbf{F})\right]\right\}$ and $\widetilde{d} = \mathbf{w}'\mathbf{e}$. Let the N eigenvalues of Ω be $\left(\widetilde{\phi}_1, \ldots, \widetilde{\phi}_N\right)$, and the corresponding eigenvectors be $\widetilde{\mathbf{w}}_1, \ldots, \widetilde{\mathbf{w}}_N$. Then an optimal $\mathbf{w}$ is the one that yields the minimum of $\left\{\frac{\widetilde{\phi}_1}{\widetilde{d}_1^2}, \ldots, \frac{\widetilde{\phi}_N}{\widetilde{d}_N^2}\right\}$. Let the solution be $\widetilde{\mathbf{w}}^l$, then the optimal predictor of y_t is equal to

$$\widehat{y}_t = a + \mathbf{f}_t'\mathbf{w}^{EIG2}, \tag{26}$$

where $\mathbf{w}^{EIG2} = \frac{1}{\widetilde{d}_l}\widetilde{\mathbf{w}}^l$, and $a = Ey_t - E\left(\mathbf{f}_t'\right)\mathbf{w}^{EIG2}$.

When y_t and $\mathbf{f}_t$ are stationary, we may approximate Ey_t and $E\mathbf{f}_t$ by their time series mean $\bar{y} = \frac{1}{T_1}\sum_{t=1}^{T_1} y_t$ and $\bar{\mathbf{f}} = \frac{1}{T_1}\sum_{t=1}^{T_1}\mathbf{f}_t$. Then $\widehat{\mathbf{w}}^{EIG2}$ can be derived from sample approximation of Ω,

$$\widehat{\Omega} = \frac{1}{T_1}\sum_{t=1}^{T_1}\left[\mathbf{e}\left(y_t - \bar{y}\right) - \left(\mathbf{f}_t - \bar{\mathbf{f}}\right)\right]\left[\left(y_t - \bar{y}\right)\mathbf{e}' - \left(\mathbf{f}_t - \bar{\mathbf{f}}\right)'\right]. \tag{27}$$

This method will be denoted as *EIG2*.

Because eigenvector approach treats y and $\mathbf{f}$ symmetrically, the performance of the eigenvector approach could be severely impaired by one or a few predictive models that produce forecasts much worse than average forecasts. Aiolfi and Timmermann (2006) suggest to sort forecasts into quartiles based on their historical forecasting performance up to the point of prediction, say s. For each quartile, a pooled forecast is then computed. If the transition probability estimates (using information up to time s) suggest that a particular quartile of models produced better than average forecasts, then the pooled forecast from models in this quartile is included in the least squares estimates of the combination weights. We follow their idea here to sort out underperforming predictive models, namely, we rank the performance of each forecast. If the relative ranking stays more or less constant over time (i.e. there is a persistence in the performance of a predictive model), we use models in the top two quartiles. If the relative ranking at each period varies, then we use models in the top three quartiles or all predictive models. The reason for combining a larger number of forecasts in the latter case is because if the relative ranking of predictive models vary over time, it could be an indication of frequent "breaks" from a modeling perspective. When there is a break, a model that performed well in the past does not necessarily mean it will perform well in the future. Averaging could be a good way to hedge against model misspecification. We will denote the weight generated by this approach *EIG3*. (or *EIG4*, respectively, depends on whether the trimmed eigenvector is based on the mean-squared prediction error matrix or covariance matrix of the trimmed forecasts).

5. Mean and scale corrected simple averaging

The eigenvector approach can yield optimal combination weight only if Σ is known or $\frac{N}{T} \to 0$. If $\frac{N}{T} = c \gg 0$, then the sample estimated eigenvector could point in a random direction (Nadler, 2008). For instance, Hsiao (2012) considers 50 observed values of 10 cross-sectional units generated from a multivariate normal distribution with a given covariance matrix. The true eigenvectors and the eigenvectors based on the sample estimate of Σ from these 50 independently observed values of 10 cross-sectional units are quite different, especially for the eigenvectors corresponding to the small eigenvalues. Hence using the estimated eigenvector may actually yield suboptimal combination in finite sample. Therefore, even though the scale-adjusted eigenvector is supposed to yield the best forecast combination in principle, it often yields suboptimal combination in practice. The issue is akin to multicollinearity issue in sample estimates. It is well known when the regressors are collinear, the least square estimates are very unstable. The coefficients can change widely with a slight change of the number of sample observations.

On the other hand, simple averaging of all predictive models could be a robust way to generate prediction in finite sample. It has been documented that simple average can produce good forecasts (e.g. Bryan and Molloy (2007), Stock and Watson (2004), Timmermann (2006)). However, some or all N predictive models could be biased as noted by Palm and Zellner (1992). We can correct for possible bias of simple average method by considering a mean corrected simple averaging (*MCSA*),

$$\widehat{y}_t = \mu + \bar{y}_t, \tag{28}$$

where $\bar{y}_t$ denotes the simple averaging predictor for y_t. The mean μ is obtained as the average of $(y_t - \bar{y}_t)$.[2]

In addition to correcting the bias by adding an intercept to the simple averaging predictor, we can also make a scale correction by considering the predictive model (*MSCSA*),

$$\widehat{y}_t = \mu + c\bar{y}_t, \tag{29}$$

where the mean μ and scale c are obtained by regressing y_t on a constant and $\bar{y}_t$.

[2] This is also the approach suggested by Capistran and Timmermann (2009). We wish to thank a referee for calling our attention to this.

5.1. Conditions for simple average being optimal

One of the puzzles for forecast combination is the documentation of simple average (or equally weighted combination) dominating more sophisticated forecast combinations (e.g. Huang and Lee (2010), Palm and Zellner (1992) and Stock and Watson (2004)). Eq. (17) says that simple average can be optimal if and only if

$$\mathbf{e} = c\Sigma^{-1}\mathbf{e}, \tag{30}$$

where c is a positive constant. Eq. (30) also implies

$$\Sigma\mathbf{e} = c\mathbf{e}. \tag{31}$$

Proposition 5.1. *When (30) (or (31)) holds, the geometric approach of finding* $\mathbf{w}$ *to minimize MSPE (1) is identical to the regression approach (Granger and Ramanathan (1984), GR1 (3)).*

Proof. We have shown in Remark 3.2 that minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ is identical to *GR*1. When (30) holds, minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ yields optimal (17)

$$\mathbf{w}^{VC} = \frac{1}{N}\mathbf{e}. \tag{32}$$

The geometric approach to choosing $\mathbf{w}$ is to find $\mathbf{w}$ corresponding to the min $\left(\frac{\phi_1}{d_1^2}, \frac{\phi_2}{d_2^2}, \ldots, \frac{\phi_N}{d_N^2}\right)$. Since Σ is positive definite, all its eigenvalues, $\lambda_j > 0$ for $j = 1, \ldots, N$. Let $\mathbf{w}_j$ be the eigenvector corresponding to

$$\Sigma\mathbf{w}_j = \phi_j\mathbf{w}_j. \tag{33}$$

When (30) holds,

$$\mathbf{e}'\Sigma\mathbf{w}_j = c\mathbf{e}'\mathbf{w}_j = \phi_j\mathbf{e}'\mathbf{w}_j, \quad j = 1, \ldots, N. \tag{34}$$

Since $\phi_j > 0$ for $j = 1, \ldots, N$, any $\phi_j \neq c$, $\mathbf{e}'\mathbf{w}_j = 0$.

When $\phi_l = c$, the corresponding eigenvector $\mathbf{w}_l$ takes the form $\mathbf{w}_l = \frac{1}{\sqrt{N}}\mathbf{e}$ and

$$\mathbf{w}_l'\Sigma\mathbf{w}_l = c = NE\left(y_t - \frac{1}{N}\mathbf{e}'\mathbf{f}_t\right)^2. \tag{35}$$

On the other hand, any other eigenvector, $\mathbf{w}_j$, will be subject to $\mathbf{e}'\mathbf{w}_j = 0$. Then

$$\mathbf{w}_j'\Sigma\mathbf{w}_j = E\left(\mathbf{w}_j'\mathbf{f}_t\right)^2$$

is not related to the *MSPE*. In other words, minimizing $\mathbf{w}'\Sigma\mathbf{w}$ subject to $\mathbf{e}'\mathbf{w} = 1$ or $\mathbf{w}'\mathbf{w} = 1$ yields identical weight which is identical to *GR*1 when (30) (or (31)) holds.

To see under what conditions (30) will be satisfied, we decompose the prediction error of the ith model, f_{it}, into two components, a component due to the impact of r common factors, $\boldsymbol{\lambda}_t$, that are omitted from all predicted models and an idiosyncratic component, ε_{it}, that is independent across i,

$$y_t - f_{it} = \mathbf{b}_i'\boldsymbol{\lambda}_t + \varepsilon_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T_1 \tag{36}$$

where $\mathbf{b}_i$ denotes the impact of $\boldsymbol{\lambda}_t$ on the error of ith predictive model. For simplicity, we assume all predictive models yield unbiased predictors (i.e., $\alpha_i = 0$) and $E\left(\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t'\right) = \mathbf{I}_r$, and $E\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right) = \mathbf{D}$ being diagonal. Then

$$\Sigma = \mathbf{B}\mathbf{B}' + \mathbf{D}. \tag{37}$$

In other words, the correlations across different predictor errors are due to the omission of common factors, $\boldsymbol{\lambda}_t$ in their models even though the impact of $\boldsymbol{\lambda}_t$ is model specific, $\mathbf{b}_i$. We assume *rank* $(B) = r < N$. If $r = N$, then Σ is unrestricted.

Lemma 5.1. *The simple average yields optimal forecast combination if and only if* $\mathbf{e} = c\Sigma^{-1}\mathbf{e}$ *(or* $\mathbf{e} = c\Sigma\mathbf{e}$*), where* $\mathbf{e}' = (1, \ldots, 1)$*. If we decompose the forecast error of the ith model as the sum of two components, a component due to the impact of* r $(r < N)$ *common factors* $\boldsymbol{\lambda}_t$ *with impact* $\mathbf{b}_i$ *and an idiosyncratic component* ε_{it} *as in (36), then a simple average yields optimal forecast combination:*

(a) *if* $\sum_{i=1}^N \mathbf{b}_i = 0$ *when* $\mathbf{D}$ *is proportional to an identity matrix, or,*
(b) $\sum_{j=1}^N \mathbf{b}_j'\mathbf{b}_i + d_i = \sum_{j=1}^N \mathbf{b}_j'\mathbf{b}_{i'} + d_{i'} = d^*$ *for all* i *and* i'.

Proof. (a) When $\mathbf{D} = \sigma^2\mathbf{I}$, then

$$\left|\mathbf{B}\mathbf{B}' + \mathbf{D} - \lambda\mathbf{I}\right| = \left|\mathbf{B}\mathbf{B}' - \left(\lambda - \sigma^2\right)\mathbf{I}\right|. \tag{38}$$

Then the eigenvalue $\lambda_i \geq \sigma^2$. When an eigenvector lies on the null space of $\mathbf{B}$, the corresponding eigenvalue $\lambda^* = \sigma^2$, which is the smallest. When $\sum_{i=1}^N \mathbf{b}_i = 0$, $\mathbf{w} = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)'$ is lying on the null space of $\mathbf{B}$.

(b) Let d^* be the smallest diagonal element of (37). When (b) is satisfied, $\mathbf{w} = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)'$ is the eigenvector to the eigenvalue d^*.

Remark 5.1. Timmermann (2006) has shown that equal weights are optimal when the individual forecast errors have the same variance, σ^2, and identical pair-wise correlations, ρ. In this case,

$$\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho)}\left[\mathbf{I} - \frac{\rho}{1 + (N-1)\rho}\mathbf{e}\mathbf{e}'\right].$$

Then

$$\Sigma^{-1}\mathbf{e} = \frac{1}{\sigma^2\left[1 + (N-1)\rho\right]}\mathbf{e}.$$

In other words, Timmermann's condition or Lemma 5.1 gives sufficient conditions for a simple average to be optimal. It is not a necessary condition. A necessary and sufficient condition for the simple average to be optimal is (30) or (31).

Remark 5.2. When conditions of Lemma 5.1 are not satisfied, the equally weighted combination, $\widetilde{\mathbf{e}}' = \frac{1}{\sqrt{N}}(1, \ldots, 1)$ is no longer optimal. However, $\widetilde{\mathbf{e}}'\Sigma\widetilde{\mathbf{e}}$ may still be smaller than $\mathbf{w}^{*\prime}\Sigma\mathbf{w}^*$, where $\mathbf{w}^*$ is the eigenvector corresponding to the smallest eigenvalue of the sample mean squared error matrix, $\mathbf{S}$. As discussed by Baik and Silverstein (2006), Nadler (2008) and Hsiao (2012, Tables 1 and 2), the sample estimates of Σ, $\mathbf{S}$, can be very different from Σ and the eigenvector can point to a random direction when $\frac{N}{T_1} = c$ does not go to 0.

6. Monte Carlo studies

In this section we conduct a small scale Monte Carlo study to evaluate the finite sample performance of our proposed forecast combination methods, the eigenvector approach (*EIG*1), the bias corrected eigenvector approach (*EIC*2), the trimmed eigenvector or trimmed mean corrected eigenvector approach using the top two quartile predictive models only (*EIG*3 or *EIG*4), a mean corrected simple average (*MCSA*) and a mean and scale corrected simple average (*MSCSA*) to some popular forecast combination methods such as the simple average – *SA* (*N*), Bates and Granger (1969), *BG*, (7), variance–covariance approach, *VC*, (17), Granger and Ramanathan (1984) models – *GR*1, *GR*2 and *GR*3 ((3), (4), (5)), Swanson and Zeng (2001) model selection approaches – *AICFC*, *AICCFC* and *BICFC* and Bayesian averaging (Buckland et al., 1997) – *MABICFC*. For a fair comparison between the geometric and regression approach, we also provide the results of regression analysis using the same set of models as in *EIG*3 or *EIG*4. Since we do know the data generating process in the experiments as benchmarks, we also provide results based on information combination methods,

TVC, *EVC*, and results based on model selection criterion *AIC*, *AICC*, *BIC* and their Bayesian averaging counterparts *MABIC*.

We consider four experimental designs. In all of the following designs, we divide the horizon of study into three periods by $T_0 = 100$, $T_1 = 200$ and $T = 220$. The sample mean squared prediction error of various methods are computed using observations y_t and predictive models $\widehat{y}_t$ from $T_1 + 1$ to T. The weight of various forecast combinations are derived from observed values y_t and the available forecasts $\mathbf{f}_{i,t}$ using data from $T_0 + 1$ to T_1. The parameters of predictive models $\mathbf{f}_{i,t}$ are estimated using data from 1 to T_0. For model selection criteria, *AIC*, *AICC*, *BIC* and their Bayesian averaging counterpart *MABIC*, the first T_1 observations are used to estimate parameters and the values of the three model selection criteria.

6.1. Design 1

The DGP of $\{y_t, x_{kt}, k = 1, \ldots, K, t = 1, \ldots, T\}$ with $K = 10$ is generated by

$$y_t = \sum_{k=1}^{K} \theta_k x_{kt} + e_t, \quad e_t \backsim N(0, 1),$$

where the first three $\mathbf{x}$ follow $AR(1)$ structure

$$x_{1t} = 0.5x_{1,t-1} + u_{1t}, \quad u_{1t} \backsim N(0, 0.75),$$
$$x_{2t} = 0.7x_{2,t-1} + u_{2t}, \quad u_{2t} \backsim N(0, 0.5),$$
$$x_{3t} = 0.9x_{3,t-1} + u_{3t}, \quad u_{3t} \backsim N(0, 1.5),$$

the fourth and fifth $\mathbf{x}$ follow $U(1, 2)$ and $x_{5t} = U(1, 4)$ respectively, the sixth and seventh follow *beta*(1, 1) and *beta*(1, 3) distribution, the eighth and ninth have *gamma*(3) and *gamma*(1) structure, the last one is normally distributed as $N(0, 0.5)$. Following Hansen (2008), we generate the coefficients of x_{kt} as

$$\theta_k = \sqrt{\frac{p}{1-p}}\gamma_k, \tag{39}$$

$$\gamma_k = \left(\frac{k}{K+1}\right)^{\alpha}\left(1 - \frac{k}{K+1}\right)^{\beta} \Bigg/ \sum_{k=1}^{K}\left[\left(\frac{k}{K+1}\right)^{\alpha}\left(1 - \frac{k}{K+1}\right)^{\beta}\right]^2, \tag{40}$$

$$= k^{\alpha+\beta}\left(\frac{K+1}{k}\right)^{\beta} \Bigg/ \sum_{k=1}^{K} k^{2(\alpha+\beta)}\left(\frac{K+1}{k}\right)^{2\beta} \tag{41}$$

where p takes on the value on the grid of $\{0.1, 0.2, \ldots, 0.9\}$ which are the third rows of Table 1.

The true covariance matrix elements are $\sigma_y^2 = \sigma_e^2 = 1$, $\sigma_{yx} = \left(\theta_1\sigma_1^2, \ldots, \theta_{10}\sigma_{10}^2\right)$, $\Sigma_{xx} = diag\left(\sigma_1^2, \ldots, \sigma_{10}^2\right)$, where σ_j^2 are the true variances of $x_j, j = 1, \ldots, 10$ respectively. σ_1^2 to σ_3^2, the variances of the *AR* processes, are equal to $\sigma_{u_j}^2/\left(1 - \rho_j^2\right)$, where ρ_j is the $AR(1)$, σ_4^2 and σ_5^2, the variances of uniform distribution $U(a, b)$, are equal to $\frac{1}{12}(b-a)^2$, σ_6^2 and σ_7^2, the variances of *beta* (a, b), are equal to $ab/[(a+b)^2(a+b+1)]$, σ_8^2 and σ_9^2, the variances of *gamma* (a), are equal to a, the variance of normally distributed variable, x_{10}, σ_{10}^2, is equal to 0.5.

There are 8 predictive models. None of these completely capture the true data generating process. The first model regresses y_t on constant and x_{1t}. The second model regresses y_t on constant and x_{1t}, x_{2t} and so on until the eighth regression which is y_t on constant and $x_{1t}, \ldots, x_{8,t}$. We consider ten different combinations of (α, β), $(-1, 1)$, $(-0.5, 1)$, $(-1, 1.25)$, $(-0.5, 1.25)$, $(-1, 1.5)$, $(-0.1, 0.8)$, $(0.3, 0.6)$, $(0.5, 0.5)$, $(0.8, 0.2)$, $(0.9, 0.1)$. The first five designs yield predictive models that are of similar magnitude of predictive accuracy. The last five designs yield predictive models of diverse accuracy (see Table 3). We therefore only report the results of design $(-1, 1)$ and $(-0.1, 0.8)$ as representative results for designs 1 to 5 or 6 to 10 in Tables 1 and 2. The leftmost column of Tables 1 and 2 present the results for $p = 0.1$, while the rightmost column the results of $p = 0.9$. (The results for other designs are available from the authors.)

6.2. Design 2

The DGP of y_t is

$$y_t = e_t + \sum_{k=1}^{q} \theta_k e_{t-k}, \quad e_t \backsim N(0, 1),$$

the coefficients of e_{t-k} follow $\theta_k = k^{-r/2}$, $k = 1, \ldots, q = 80$, $r = 1.2, 1.4, \ldots, 3$. This ensures that $\Sigma_k \theta_k^2 < \infty$.

The 10 predictive models are $AR(k)$, $k = 1, \ldots, 10$ with a constant.

The true covariance matrix elements are $\sigma_y^2 = \left(1 + \sum_{k=1}^{q} \theta_k^2\right)\sigma_e^2$. $\Sigma_{xx} = \sigma_e^2 \mathbf{I}_{q+1}$.

$\sigma_{yx} = \sigma_e^2 \boldsymbol{\theta}$, where $\boldsymbol{\theta} = \left(1, \theta_1, \ldots, \theta_q\right)$.

6.3. Design 3

Same as Design 1 except $\{f_t(m), m = 1, \ldots, 8\}$ are generated without constant terms, i.e. predictive models are biased. The procedure is regressing y_t on $\{x_{1t}\}$ for model 1, y_t on $\{x_{1t}, x_{2t}\}$ for model 2, and y_t on $\{x_{1t}, x_{2t}, \ldots, x_{8t}\}$ for model 8. The true covariance matrix is the same as in Design 1.

6.4. Design 4

We suppose that $(y, \mathbf{x})$ have a common factor structure, where $\mathbf{x}$ is a 10×1 vector.

$$\begin{pmatrix} y_t \\ \mathbf{x}_t \end{pmatrix} = \mathbf{B}\boldsymbol{\lambda}_t + e, e \backsim N(0, 1).$$

DGP of the factors are

$$\lambda_{1t} = 0.5\lambda_{1,t-1} + u_{1t}, \quad u_{1t} \backsim N(0, 0.75),$$
$$\lambda_{2t} = 0.7\lambda_{2,t-1} + u_{2t}, \quad u_{2t} \backsim N(0, 0.5),$$
$$\lambda_{3t} = 0.9\lambda_{3,t-1} + u_{3t}, \quad u_{3t} \backsim N(0, 1.5).$$

Factor loadings $\mathbf{b}_i$ are generated by $\alpha\mathbf{1} + N\left(0, \beta^2\mathbf{I}_3\right)$ and then adjusted by the row sum-square. For example, when $N = 2$ and $r = 2$, then the loading matrix is

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}' \\ \mathbf{B}_x' \end{pmatrix} \begin{pmatrix} b_{11}/\left(b_{11}^2 + b_{12}^2 + b_{13}^2\right) & b_{21}/\left(b_{21}^2 + b_{22}^2 + b_{23}^2\right) \\ b_{12}/\left(b_{11}^2 + b_{12}^2 + b_{13}^2\right) & b_{22}/\left(b_{21}^2 + b_{22}^2 + b_{23}^2\right) \\ b_{13}/\left(b_{11}^2 + b_{12}^2 + b_{13}^2\right) & b_{23}/\left(b_{21}^2 + b_{22}^2 + b_{23}^2\right) \end{pmatrix}.$$

The 10 predictive models are constructed through regression of y_t on $\{1, x_{1t}\}$, regression of y_t on $\{1, x_{1t}, x_{2t}\}$, until regression of y_t on $\left\{1, x_{1t}, \ldots, x_{10,t}\right\}$. The true covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} = \begin{pmatrix} \mathbf{b}'\mathbf{b} + 1 & \mathbf{b}'\mathbf{B}_x \\ \mathbf{B}_x'\mathbf{b} & \mathbf{B}_x'\mathbf{B}_x + \mathbf{I}_N \end{pmatrix}. \tag{42}$$

Table 8 reports the results.

6.5. Results

The results of Design 2–4 are similar to those reported in Tables 1 and 2 and are available from the authors. The main findings are:

Table 1
Mean squared prediction error for design 1.1 ($\alpha = -1, \beta = 1$).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TVC	1.0093	1.0084	1.003	1.009	0.9958	1.0191	0.9788	1.0027	0.9837
EVC	1.0647	1.0685	1.061	1.0569	1.0513	1.0716	1.0404	1.0543	1.038
AIC	1.0308	1.033	1.03	1.0282	1.0221	1.048	1.0129	1.0388	1.0212
AICC	1.0299	1.0317	1.0296	1.0275	1.0206	1.047	1.0118	1.0392	1.0209
BIC	1.0215	1.0217	1.0168	1.0199	1.0119	1.037	1.0037	1.0325	1.0262
MAABIC	1.0213	1.0218	1.0161	1.0192	1.011	1.0353	1.0012	1.0281	1.0175
AICFC	1.08	1.0861	1.0766	1.0738	1.0738	1.0853	1.0575	1.0872	1.0679
AICCFC	1.0793	1.0827	1.0731	1.0708	1.0716	1.083	1.0551	1.0857	1.065
BICFC	1.055	1.0607	1.0527	1.0504	1.0497	1.0676	1.0354	1.064	1.0503
MABICFC	1.0527	1.0568	1.0492	1.0473	1.0464	1.0633	1.0303	1.0589	1.0449
SA(N)	1.0608	1.0615	1.0602	1.0565	1.0478	1.0726	1.0316	1.0536	1.0333
BG	1.0593	1.0599	1.0586	1.0552	1.0466	1.0712	1.0304	1.0524	1.0324
EIG1	1.0617	1.0625	1.0612	1.0573	1.0485	1.0735	1.0323	1.0544	1.034
EIG2	1.0625	1.0618	1.0569	1.0571	1.0479	1.0748	1.0299	1.0508	1.0351
VC	1.1119	1.115	1.1054	1.1	1.094	1.1126	1.0824	1.0996	1.0735
GR1	1.1119	1.115	1.1054	1.1	1.094	1.1126	1.0824	1.0996	1.0735
GR2	1.1072	1.1111	1.1028	1.0951	1.0944	1.1119	1.0846	1.1028	1.0811
GR3	1.1185	1.1273	1.1129	1.107	1.1036	1.1208	1.0909	1.109	1.0902
MSCSA	1.0296	1.0327	1.025	1.0324	1.0259	1.0529	1.0176	1.0436	1.0355
MCSA	1.0618	1.061	1.0561	1.0564	1.0472	1.074	1.0293	1.0501	1.0345
EIG3	1.048	1.0479	1.0458	1.0439	1.0347	1.0614	1.0221	1.0449	1.0287
EIG4	1.0489	1.0476	1.0423	1.0434	1.0347	1.0634	1.0204	1.0423	1.0297
GR1T	1.0623	1.0626	1.0585	1.0567	1.0531	1.0716	1.037	1.058	1.0442
GR2T	1.0538	1.0562	1.0547	1.0473	1.0495	1.0667	1.0365	1.0591	1.0478
GR3T	1.0648	1.0694	1.064	1.0589	1.0598	1.0745	1.0409	1.0636	1.0517

Table 2
Mean squared prediction error for design 1.6 ($\alpha = -0.1, \beta = 0.8$).

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TVC	0.9978	0.9999	0.993	1.0016	0.9746	1.0007	0.9944	1.0061	0.9744
EVC	1.0541	1.0599	1.0531	1.061	1.0258	1.055	1.0492	1.0697	1.0265
AIC	1.0424	1.0451	1.0472	1.0628	1.0276	1.0629	1.0523	1.0781	1.0525
AICC	1.0413	1.0446	1.0466	1.0632	1.0284	1.0651	1.0537	1.0787	1.0525
BIC	1.052	1.0449	1.0401	1.0669	1.0376	1.0898	1.0947	1.1088	1.0541
MAABIC	1.0432	1.0424	1.0375	1.0617	1.0293	1.0789	1.0816	1.1013	1.0549
AICFC	1.0951	1.0936	1.0909	1.1161	1.0801	1.1162	1.1068	1.1308	1.1123
AICCFC	1.0935	1.0913	1.0883	1.1162	1.0778	1.1155	1.1067	1.1307	1.1132
BICFC	1.0747	1.0774	1.0779	1.1022	1.0685	1.1149	1.1119	1.1384	1.1099
MABICFC	1.0681	1.0675	1.0661	1.0899	1.0536	1.0951	1.0938	1.1199	1.0939
SA(N)	1.0528	1.0647	1.0708	1.1005	1.0775	1.1336	1.1802	1.2932	1.5679
BG	1.0523	1.0639	1.0681	1.0941	1.0624	1.1096	1.125	1.1786	1.2586
EIG1	1.0534	1.0656	1.0726	1.1038	1.0868	1.1513	1.2451	1.4685	2.0086
EIG2	1.0556	1.0646	1.0694	1.0968	1.0798	1.1471	1.2151	1.4022	2.1174
VC	1.1051	1.1036	1.105	1.1124	1.0706	1.1084	1.0945	1.1241	1.1026
GR1	1.1051	1.1036	1.105	1.1124	1.0706	1.1084	1.0945	1.1241	1.1026
GR2	1.1131	1.1098	1.1091	1.1191	1.0804	1.1125	1.1016	1.126	1.107
GR3	1.1196	1.1104	1.1082	1.1183	1.082	1.1169	1.1065	1.129	1.1136
MSCSA	1.0597	1.0712	1.0722	1.0973	1.0657	1.113	1.14	1.1963	1.3498
MCSA	1.0552	1.0642	1.0684	1.0951	1.0739	1.1339	1.1781	1.2804	1.567
EIG3	1.0542	1.0639	1.0694	1.09	1.0522	1.1009	1.1098	1.1573	1.2143
EIG4	1.0561	1.0631	1.0684	1.0886	1.0525	1.0974	1.1102	1.1556	1.2158
GR1T	1.069	1.0827	1.0874	1.1017	1.0652	1.1082	1.1005	1.1335	1.1131
GR2T	1.0729	1.0881	1.0858	1.1006	1.0681	1.1052	1.1005	1.129	1.1134
GR3T	1.0808	1.087	1.0851	1.1018	1.0649	1.1027	1.1023	1.1281	1.1128

Table 3
R^2 for the 8 predictive models for the 10 cases in design 1.

Design	Model									
	1	2	3	4	5	6	7	8	9	10
1	0.0544	0.1594	0.0392	0.1198	0.0321	0.1664	0.1013	0.0736	0.04	0.0382
2	0.0733	0.2286	0.0548	0.1752	0.0468	0.2882	0.2131	0.179	0.1264	0.1118
3	0.1031	0.4236	0.077	0.3293	0.0655	0.7634	0.7951	0.7507	0.6568	0.6133
4	0.1124	0.4313	0.0864	0.3384	0.0751	0.7711	0.8051	0.7627	0.6702	0.6267
5	0.1221	0.4451	0.0957	0.3519	0.0847	0.8051	0.8667	0.8373	0.7606	0.7244
6	0.1321	0.4518	0.1048	0.3598	0.0941	0.8097	0.8731	0.8457	0.772	0.737
7	0.1422	0.4581	0.1142	0.3667	0.1044	0.8121	0.876	0.8495	0.7776	0.7437
8	0.1524	0.4687	0.1239	0.3767	0.1144	0.8419	0.9525	0.963	0.961	0.9558

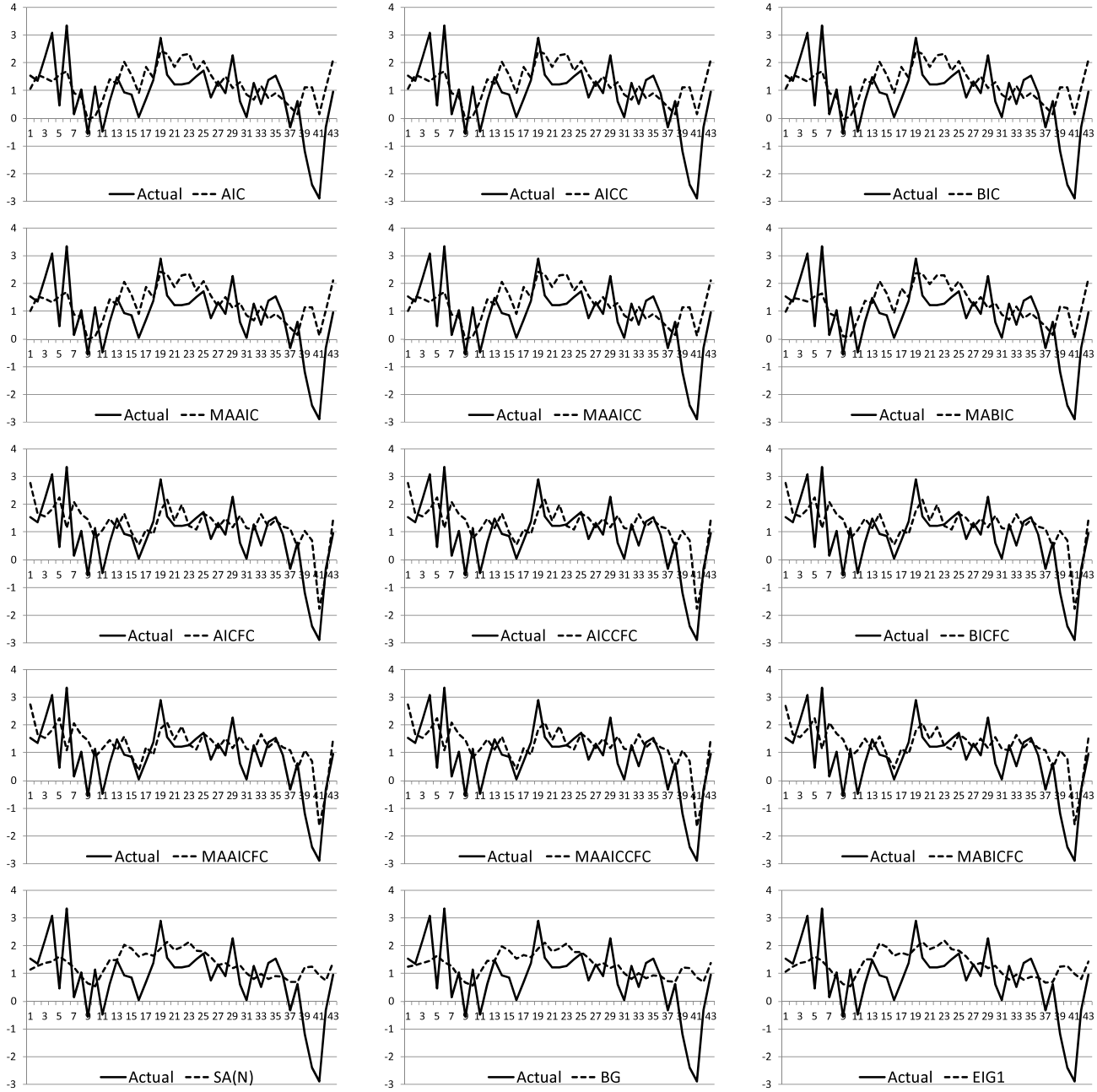

Fig. 2. Actual and predicted US real GDP under continuously updating forecasting framework.

1. When information generating the predictive models is readily available, directly combining the information optimally yields more accurate predictions than optimally combining forecasts. However, in the case that the parameters of the data generating process have to be estimated, the model averaging approach appears to dominate the model selection approach.
2. In finite sample, when information on generating the predictive models is unknown, forecast combinations appear to dominate the approach of selecting a single predictive model in terms of some model selection criterion to generate predictions.
3. When there is no structural change, simple averaging predictions do not yield more accurate prediction than optimally combining them based on some optimality criteria.
4. When the performance of all predictive models are of roughly the same magnitude, the eigenvector approach, (*EIG*1, *EIG*2) of finding the relative weight for each predictive models appears to dominate the regression approach. On the other hand, if the performance of one or more predictive models is much worse than the others, then the regression approach appears to do better (Design 1.1 vs Design 1.6, see Tables 1 and 2) since the eigenvector approach treats all predictive models symmetrically, while the regression approach will give less weight to the models which perform badly and more to the good models. Comparing the performance between *EIG*1 (or *EIG*2) vs *EIG*3 (or *EIG*4) or *GR*1, 2, 3 vs *GR*1*T*, 2*T*, 3*T* confirms the conjecture that when all predictive models are more or less in the same ballpark (Design 1.1), there is not much improvement in trimming. On the other hand, when some predictive models perform substantially worse than some other predictive models (Design 1.6), trimming leads to a substantial improvement in the eigenvector approach but not much so for the regression approach. The performance of *EIG*3 leads to a39.55% improvement for $p = 0.9$ and

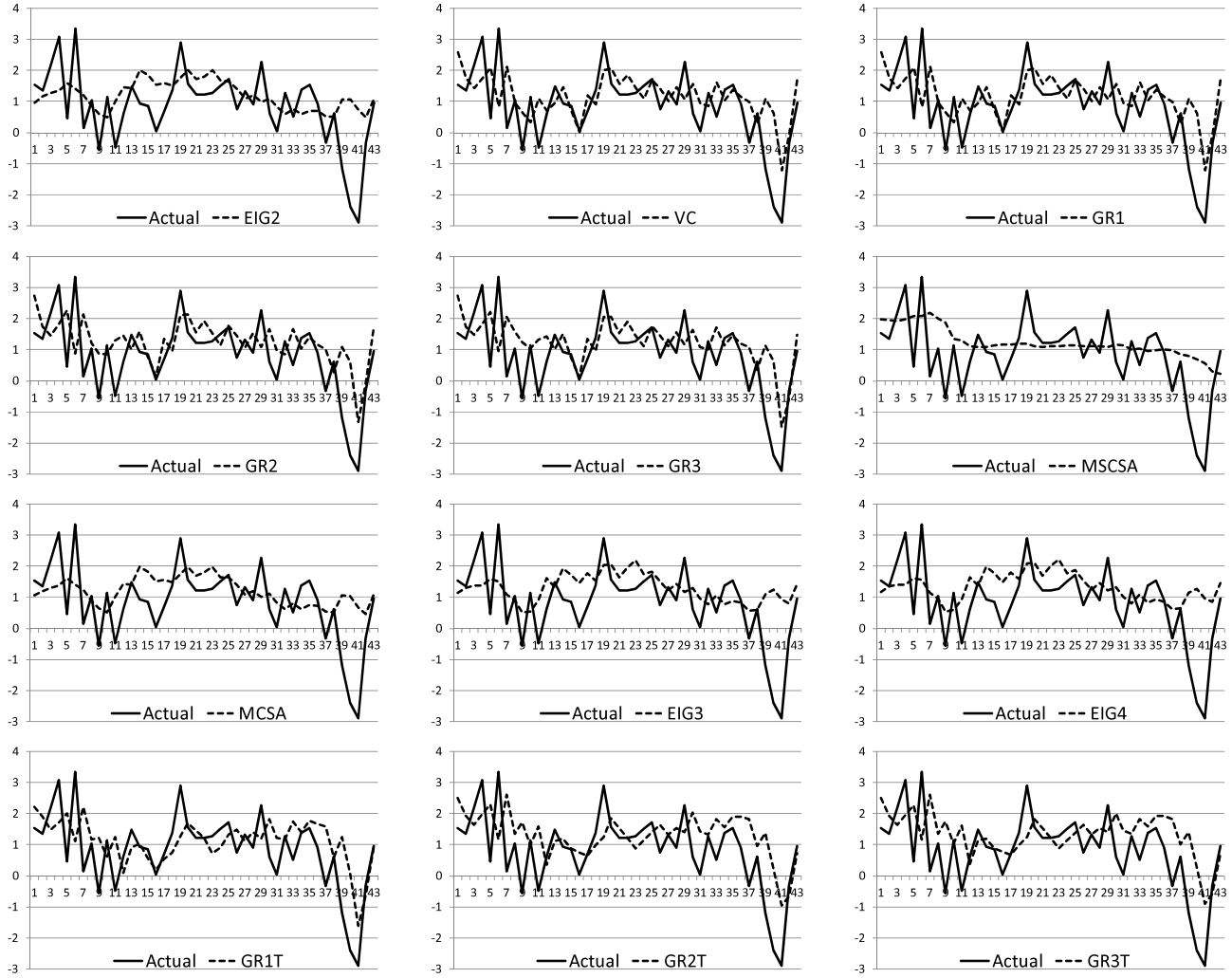

Fig. 2. (*continued*)

a 21.2% improvement for $p = 0.8$. On the other hand, there is either no improvement or it leads to worse outcomes in trimming for the regression approach. Similar conclusions are obtained for other designs. These results suggest that if there is persistence in the performance of predictive models, it is advisable to trim the models that appear to perform much worse than the average of other predictive models before applying the eigenvector approach.

5. If some predictive models are severely biased, the mean corrected eigenvector approach (*EIG*2) dominates the simple eigenvector approach (*EIG*1). (Design 1.1, 1.4, 1.7, 3.2, 3.4 and 3.6.) Otherwise, the difference between the two approaches is not that significant.
6. The normalization condition for the method of minimizing the mean squared prediction error in terms of a linear combination of errors of predictive models is not an innocuous condition. The conventional normalization condition $\sum_{i=1}^{N} w_i = 1$ in the forecast combination (Timmermann (2006) or the mean squared portfolio optimization procedure (e.g. Markowitz (1952, 1959))) actually transforms the geometric approach back to the regression approach. The Monte Carlo shows that *VC* yields identical results as *GR*1 as shown by (21) in Remark 2.3.
7. In finite sample, using less but critical sample information could yield better results than using all sample information (*BG* vs *EIG*1, *EIG*2, *GR*1, *GR*2, *GR*3).
8. Simple average (*SA* (*N*)) does not do better than combination methods based on some optimality condition. However, a mean or a mean and scale corrected simple average appears to be a robust way to combine forecasts. It yields predictions that perform well in a variety of sample designs except those cases when p is large.
9. Although one can always find a sampling approach to dominate the Bayesian approach, the optimal sampling approach of combining forecasts depends on the data generating process. On the other hand, the Bayesian approach of averaging forecasts, although not yielding optimal forecasts, appears fairly stable. The Bayesian averaging appears to yield forecasts that are always close to the optimal irrespective of the underlying data generating process.

7. Empirical applications

In this section, we examine the performance of various forecast combination methods to forecast the US real output growth rates and excess premium on S&P 500. Since we have a set of independent variables to form forecasts, we will also include model selection and Bayesian averaging approaches.

Under the fixed forecasting method, the first T_0 observations are used to calculate the parameters of our predictive models. Forecasts from $T_0 + 1$ to T are then formed. We use observations from $T_0 + 1$ to T to form forecast combination weights. Based on these weighting schemes, we combine the predicted values to form

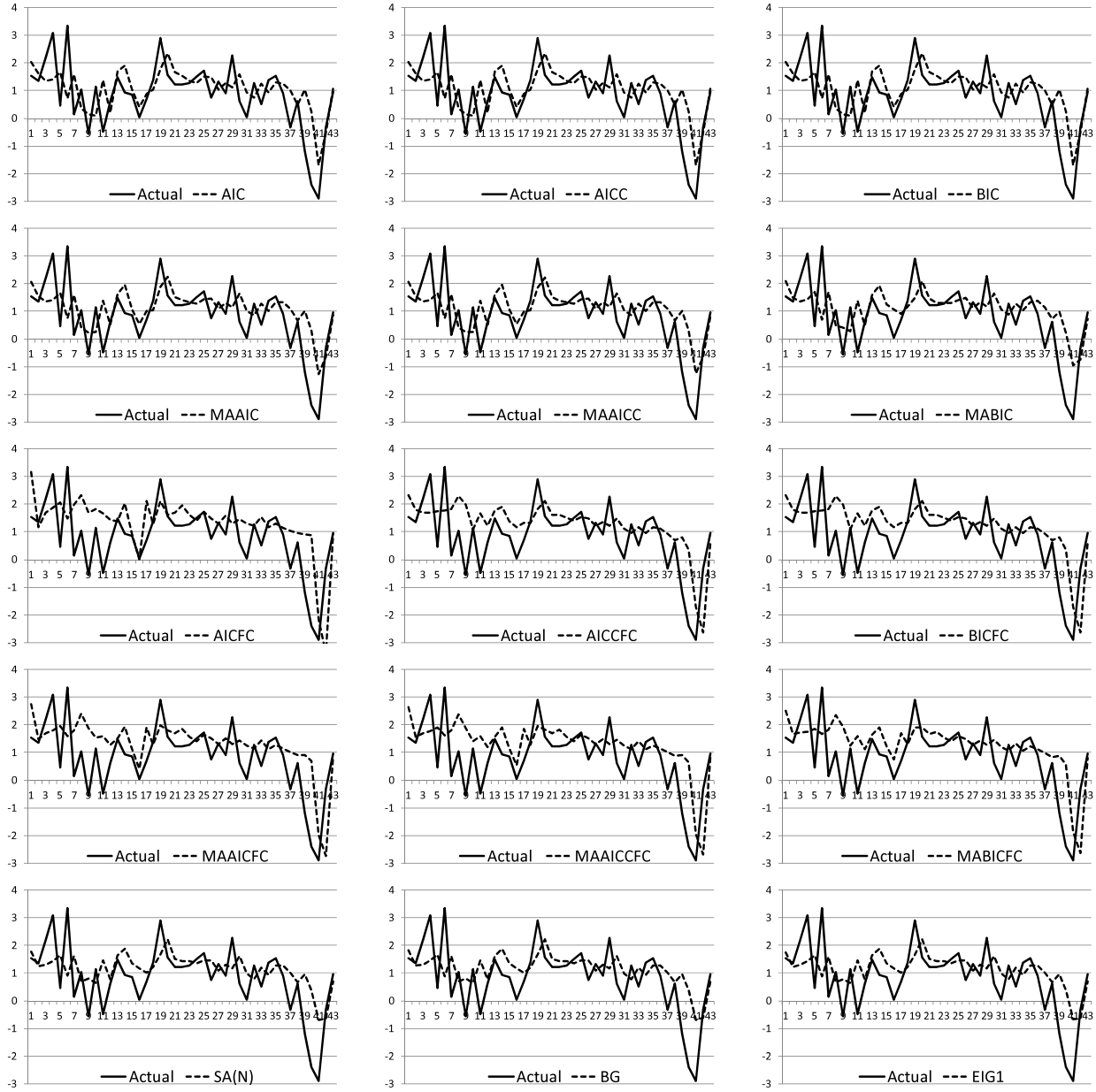

Fig. 3. Actual and predicted US real GDP under rolling forecasting framework.

a new prediction for the period from T_1+1 to T. *MSPE*s are all based on those forecasts from T_1+1 to T for fair comparison.

In addition to the fixed window forecast, we also consider the approach of a continuous updating scheme where the newest time series observations are added for estimating the parameters and for updating the combination weights. For example, the first set of regressions uses time series observations from 1 to T_0 to form forecasts at T_0+1; the second set of regressions uses observations from 1 to T_0+1 to form forecasts at T_0+2. This procedures continues until forecasts at T are obtained. The combination weights are derived based on actual y_t and predicted y_t from T_0+1 to T_1.

To hedge against possible breaks with unknown break points, we also consider a rolling window approach where we keep the window size at T_1-T_0. As time moves on, the beginning observation is dropped when a new observation becomes available for parameter estimation. The combination weights are estimated in the same way based on the actual and predicted y_t from T_0+1 to $T-1$. The first window uses y_t and predicted values of y_t from T_0+1 to T_1. The second window uses y_t and predicted values of y_t from T_0+2 to T_1+1. This procedure is carried out until combined prediction of y_T is derived.

7.1. Predicting real output growth

The predictors for the US output growth rate consist of the first lag of the dependent variable, $y_{t-1} = 400\log(GDP_{t-1}/GDP_{t-2})$, the growth rate of real GDP in quarter $t-1$; the first lag of the term spread defined as $TS_{t-1} = GB_{t-1} - FFR_{t-1}$ where GB_{t-1} is the long-term government bond yield at $t-1$ and FFR_{t-1} is the Fed Fund Rate at $t-1$; the change of the Treasury-Bill rate at $t-1$, ΔTB_{t-1}; the rate of change of seasonally adjusted $M2$, RM_{t-1}; the rate of change of S&P Industrials, RSP_{t-1}. That is, $\mathbf{x}_t = \{y_{t-1}, TS_{t-1}, \Delta TB_{t-1}, RM_{t-1}, RSP_{t-1}\}$. All the data, except real output, are obtained from International Financial Statistics, while the real output are downloaded from Federal Reserve Bank of St. Louis. Data ranges from 1970Q3 to 2009Q3, i.e. $T = 157$. We estimate the parameters with all the time series observation up to 1990Q4, i.e. $T_0 = 82$. These parameters are used to form the forecasting paths starting from 1991Q1

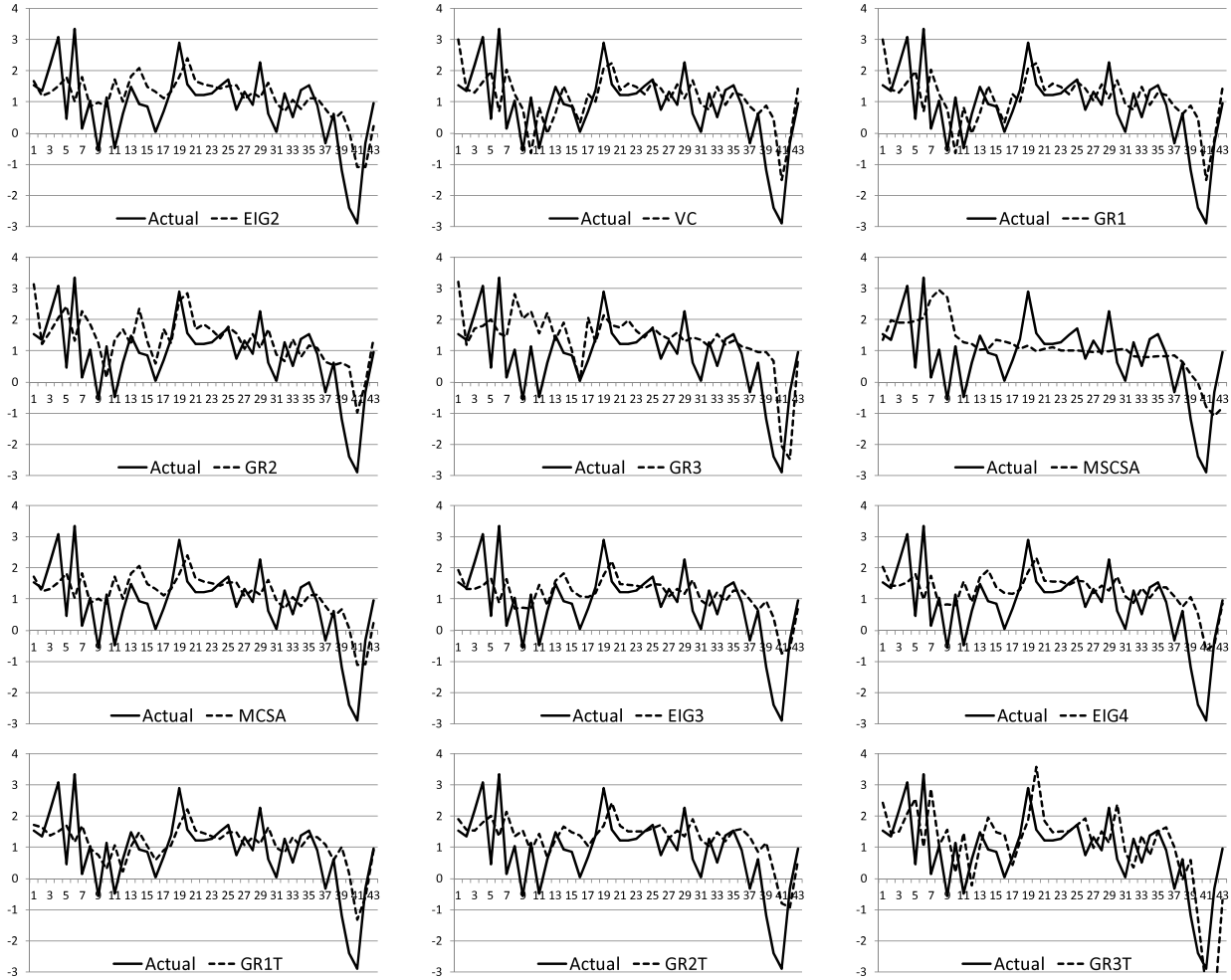

Fig. 3. (*continued*)

to 1998Q4, i.e.. $T_1 = 114$. Prediction comparisons are based on data from 1991Q1 to 2009Q3. Results are reported in Table 4. Actual and predicted path of each methods under the 2 forecasting frameworks, continuously updating and rolling window, are plotted in Figs. 2 to 3.

We note first that there are periods where the predictions based on fixed and continuously updating forecasts are way off from reality, but the rolling window approach appears able to narrow the gap. In general, the rolling framework performs better than the other 2 frameworks (except the three *GR* approaches). Second, because the information for generating the predictive model is readily available, the rolling window model selection approach of selecting the best predictive model appears to yield the most accurate predictions. Third, if information is not readily available, then the ranking of forecast combination methods in the rolling window framework appears to be consistent with the simulation results in which the eigenvector approach of obtaining relative weights for forecasting models yield more accurate predictions than the regression approach or Bayesian averaging. However, the mean corrected eigenvector approach appears to dominate the simple eigenvector approach, perhaps because some predictive models are biased. Fourth, perhaps because of frequent "breaks" between the actual and predictive models, trimming (*EIG*3, 4 or *GR*1*T*, 2*T*, 3*T*) does not lead to the improvement and the mean corrected simple average yields as good (or slightly better) forecasts as the mean corrected eigenvector approach.

7.2. Predicting excess equity premium

To predict the annual excess equity premium over S&P 500 index, which is defined as $(P_t + D_t)/P_{t-12} - 1 - TB_{t-1}$, where P_t is the closing index price on the last trading day of month t obtained in CRSP; D_t is the corresponding dividend; TB_{t-1} stands for one-month lagged US Treasury Bill rate. The nine predictors are (1) dividend yield $DY_t = \log D_t - \log P_{t-1}$, (2) one-month and (3) two-month lagged T-bill rates TB_{t-1} and TB_{t-2}, (4) rate of change of seasonally adjusted $M2$, $\Delta M_t = M_t/M_{t-12} - 1$, (5) two-month lagged inflation rate Π_{t-2}, which is computed using producer price index, (6) rate of seasonally adjusted industrial production $\Delta IP_t = IP_t/IP_{t-12} - 1$, (7) earnings price ratio $EP_t = \log E_t - \log P_t$, where E_t is 12-month moving sums of earnings on the S&P 500 index, and (8) the one-month and (9) two-month lagged government bond rate GB_{t-1} and GB_{t-2}. Except DY_t and EP_t which are obtained from Goyal and Welch (2008), other explanatory variables are obtained from International Financial Statistics.

The monthly data runs from 1960*M*3 to 2008*M*12, with total time series observations $T = 586$. We use the first 216 (1 to T_0) observations for parameter estimation and forecast, which corresponds to 1978*M*2. The forecasting paths between $T_1 = 376$, corresponding to 1991*M*6 and $T_1 - T_0$ are used to estimate the prediction error covariance matrix. The same procedures as in the above subsection are carried out. Results are reported in Table 5 and the 2 figures under various kinds of forecasting frameworks, continuously updating and rolling window, are plotted in Figs. 4 to 5.

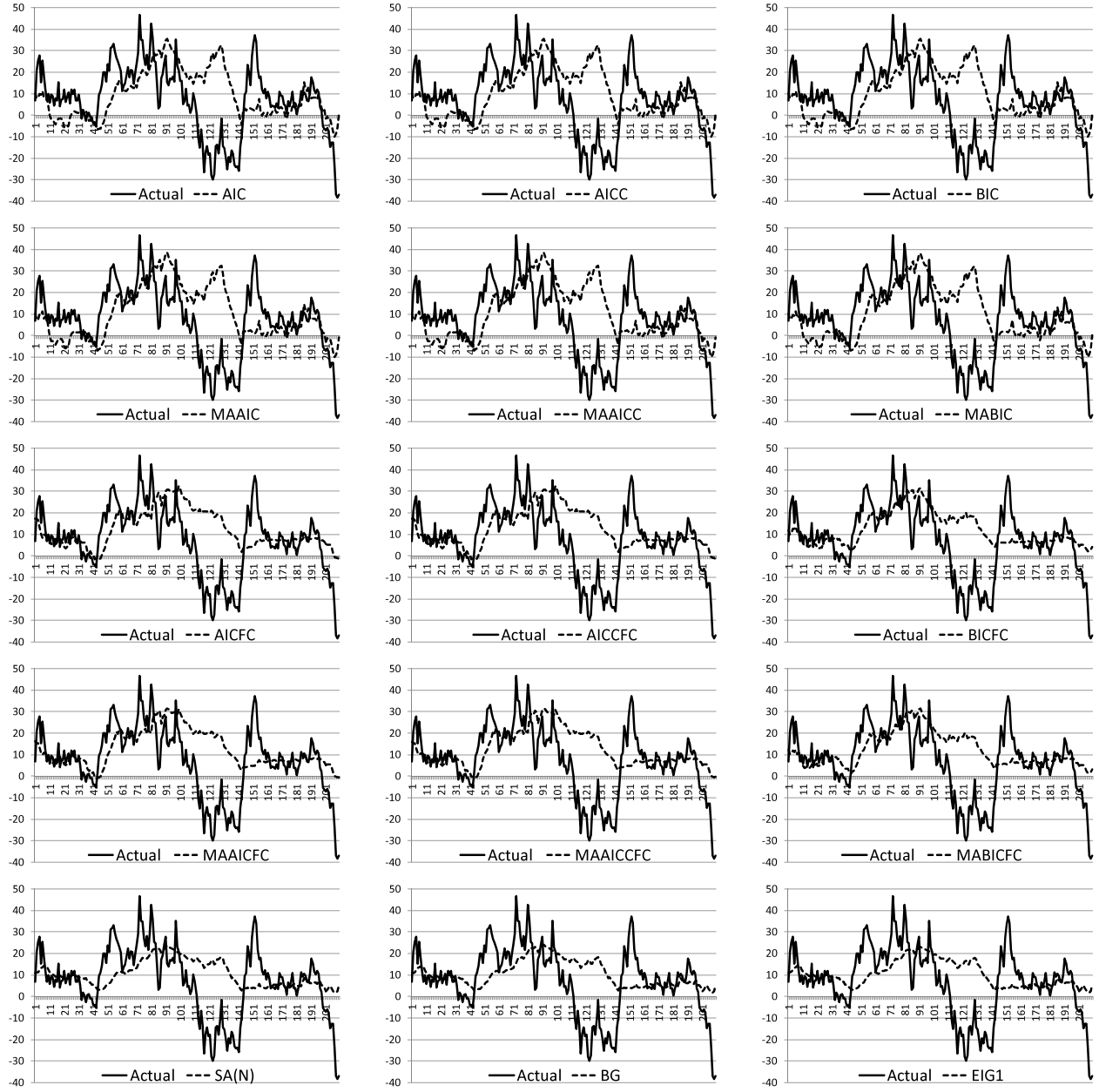

Fig. 4. Actual and predicted excess equity premium on S&P500 under recursive framework.

Again, there are periods where predictions based on fixed window or continuously updating are way off from reality but the rolling window approach is able to narrow the gaps between predictions and reality. The continuous updating and rolling window approach dominate fixed window. But now, the Bayesian averaging methods of combining the weights yield more accurate predictions than the eigenvector approach and the regression approach. However, the mean corrected simple average appears to yield predictions that are fairly robust in terms of mean squared prediction errors.

8. Concluding remarks

In this paper we have suggested several geometric approaches to combine forecasts, an eigenvector approach, a mean corrected and trimmed eigenvector approach to minimize the mean squared prediction error when there is no structural change. In deriving the optimal weight for combining the prediction error of each predictive model to yield smallest mean squared (prediction) error, we showed that the solution is not independent of the normalization condition. The conventional normalization condition that $\sum_{i=1}^{N} w_i = 1$ actually yields suboptimal solution to the normalization condition $\sum_{i=1}^{N} w_i^2 = 1$ (e.g. Timmermann (2006)) or the mean–variance portfolio allocation (e.g. Markowitz (1952, 1959)). Moreover, the normalization condition $\sum_{i=1}^{N} w_i = 1$ actually turns the problem back to the regression approach to find the optimal weight that yields identical results as Granger and Ramanathan (1984), *GR*1. We have also suggested a trimmed eigenvector approach that incorporates the clustering idea of Aiolfi and Timmermann (2006) to improve the performance of the eigenvector approach if there is persistence in the performance of predictive models.[3]

[3] An alternative to group predictive models is to follow the suggestion of Granger and Jeon (2004) to keep all close specifications, say, through some pre-test procedures.

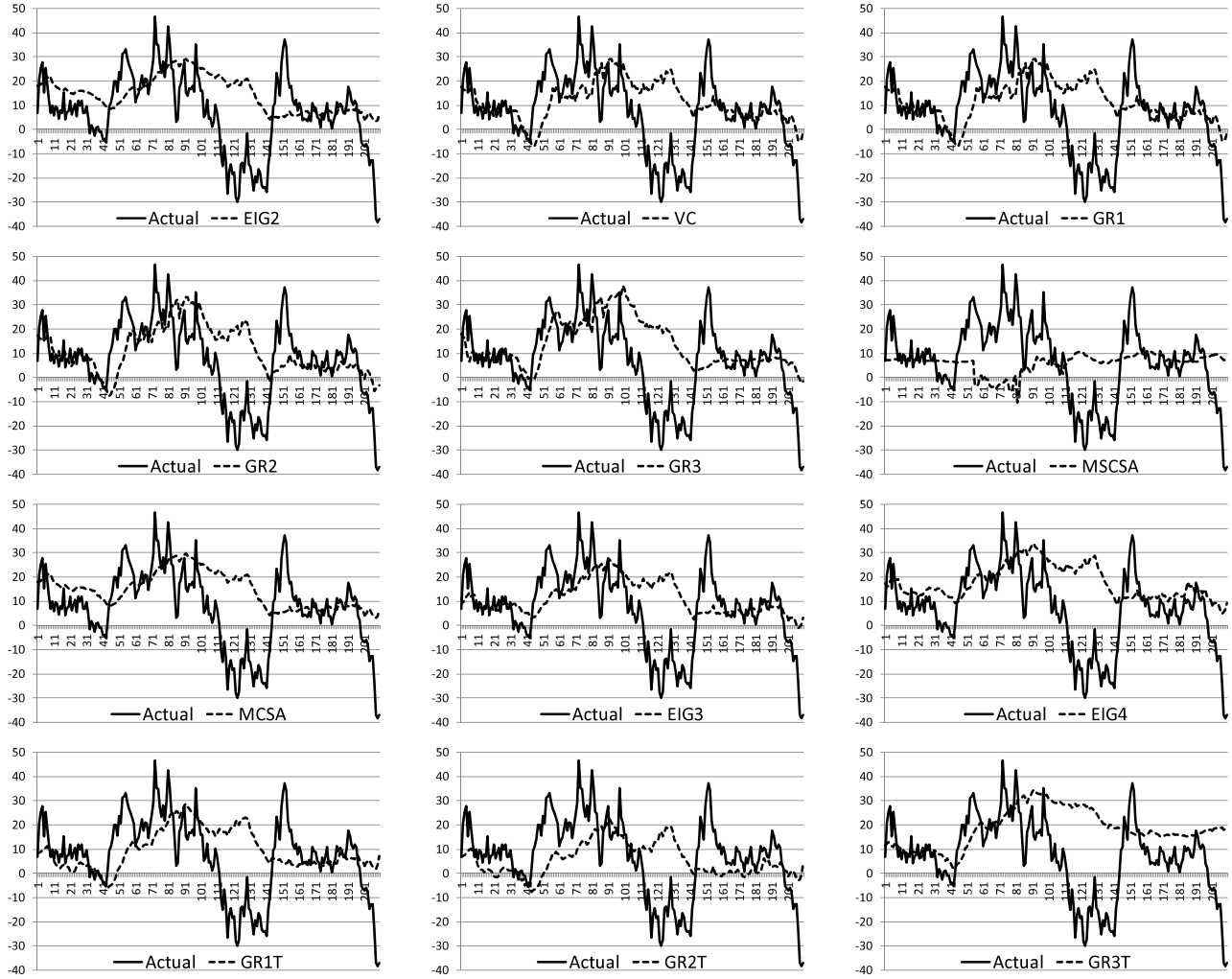

Fig. 4. (*continued*)

Table 4
MSPE of US real GDP prediction ($T_0 = 1990Q4$, $T_1 = 1998Q4$, $T = 2009Q3$).

	Fixed	Recursive	Rolling
AIC	1.4705	1.2855	1.0182
AICC	1.4705	1.2855	1.0182
BIC	1.4705	1.2855	1.0182
MABIC	1.4743	1.3016	1.1132
AICFC	1.5981	1.1777	1.476
AICCFC	1.5981	1.1777	1.2191
BICFC	1.5981	1.1777	1.2191
MABICFC	1.5799	1.2011	1.24
SA(N)	2.2513	1.4939	1.1164
BG	2.2418	1.4568	1.1161
EIG1	4.0035	1.5218	1.1175
EIG2	1.7339	1.3749	1.0936
VC	1.3683	1.1406	1.1572
GR1	1.3683	1.1406	1.1572
GR2	1.4847	1.1917	1.2654
GR3	1.6512	1.2019	1.4307
MSCSA	2.0692	1.3529	1.3522
MCSA	1.7301	1.3574	1.0916
EIG3	2.1475	1.4243	1.0995
EIG4	1.6791	1.4587	1.1752
GR1T	2.8641	1.1682	0.9562
GR2T	1.2434	1.4344	1.2357
GR3T	1.5082	1.4516	1.6799

Table 5
MSPE of predicting excess equity premium on S&P500 ($T_0 = 1990M12$, $T_1 = 1998M12$, $T = 2008M12$).

	Fixed	Recursive	Rolling
AIC	477.9	342.5	252.9
AICC	477.9	342.5	252.9
BIC	477.9	342.5	252.9
MABIC	476.7	339.7	235.0
AICFC	1414.9	291.2	191.6
AICCFC	1414.9	291.2	196.0
BICFC	1314.0	262.4	192.2
MABICFC	1260.9	266.1	190.7
SA(N)	563.2	266.5	232.2
BG	509.5	266.1	230.4
EIG1	567.4	266.4	233.9
EIG2	1074.1	303.5	225.0
VC	336.6	295.5	227.8
GR1	336.6	295.5	227.8
GR2	354.0	280.5	191.1
GR3	1453.7	302.0	191.6
MSCSA	627.5	329.6	360.5
MCSA	1043.0	303.4	222.7
EIG3	512.3	273.0	219.2
EIG4	544.7	365.3	291.6
GR1T	443.2	309.1	205.0
GR2T	387.6	286.6	186.4
GR3T	824.3	452.9	234.7

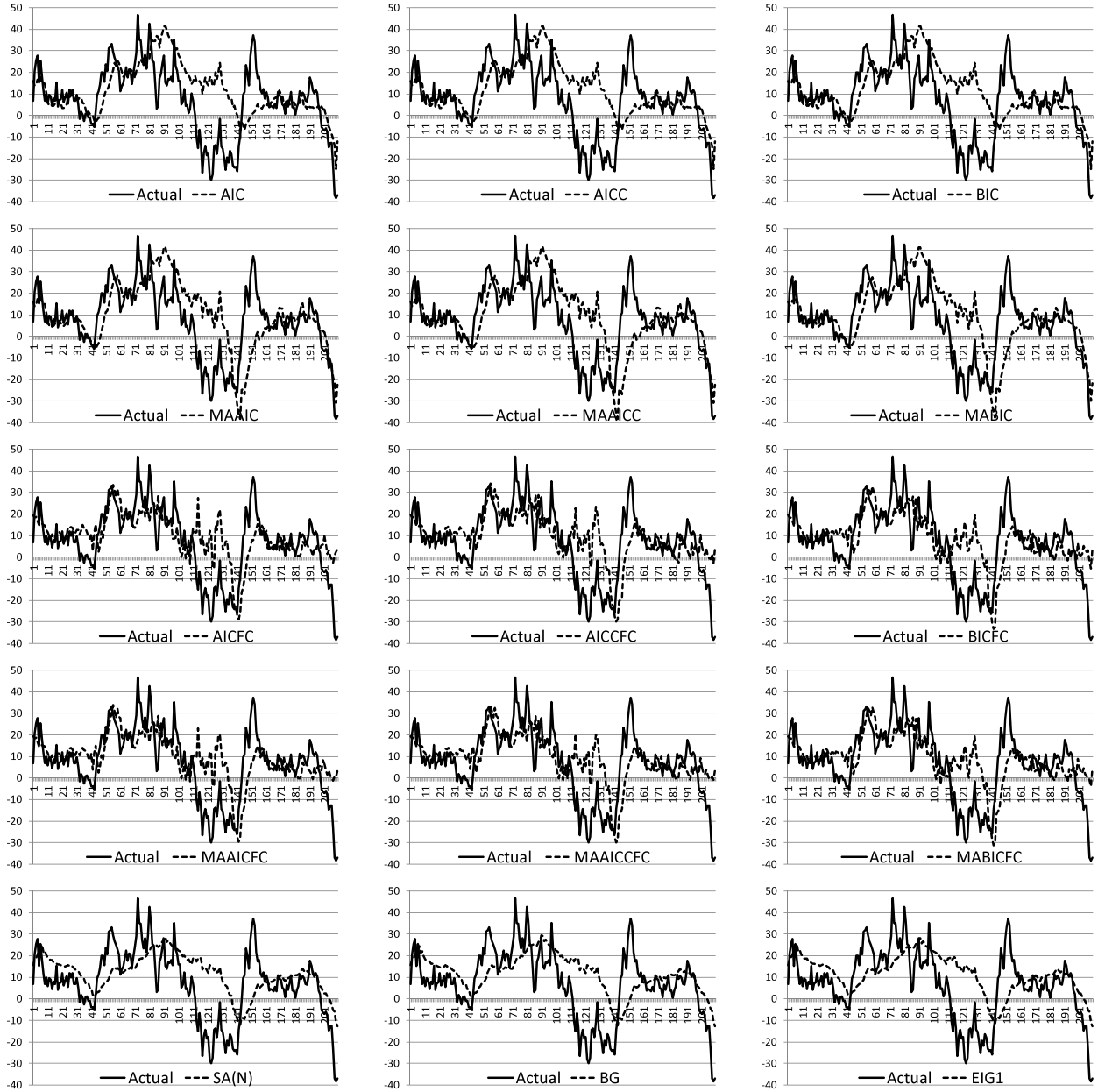

Fig. 5. Actual and predicted excess equity premium on S&P500 under rolling forecasting framework.

We have also considered mean corrected and mean and scale corrected simple average methods as a robust alternative to combine forecasts when sample size is finite and there could be structural changes. We have provided conditions where simple average could yield optimal combinations and showed that under such conditions the geometric approach and regression approach yield identical combination rules.

We have conducted Monte Carlo studies to compare the finite sample performance of the forecasting combination methods vis-a-vis some popular combination methods and information combination methods. We find that if information is readily available, then information combination is always preferred to forecast combination. However, if information is not readily available, then forecast combination based on some optimality criterion is preferred to model selection or simple averaging of all predictive models. It also appears that the eigenvector approach of combining forecasts is preferred to the regression approach when the performances of different predictive models are roughly in the same ballpark. However, if one or more predictive models perform substantially better than the others, then the regression approach is preferred and the trimmed eigenvector approach dominates the simple eigenvector approach. As a matter of fact, in a finite sample, using less but critical information could be better than using all sample information as demonstrated by comparing results based on the Bates and Granger (1969) combination method vs eigenvector or Granger and Ramanathan (1984) regression approach. The mean corrected or mean and scale corrected simple average also appears to be a robust alternative.

We also applied various forecast combination methods to predict the U.S. real GDP growth and excess equity premium. Given that there could be structural breaks from predictive models perspective, we also suggest a rolling window approach to combine forecasts. The rolling window approach dominates the fixed window or continuously updating approach. Moreover, the rolling window mean corrected or mean and scale corrected simple average actually yield more accurate forecasts than all the forecast combination methods. This is probably because in addition to the fact that there could be structure breaks, our forecast models are

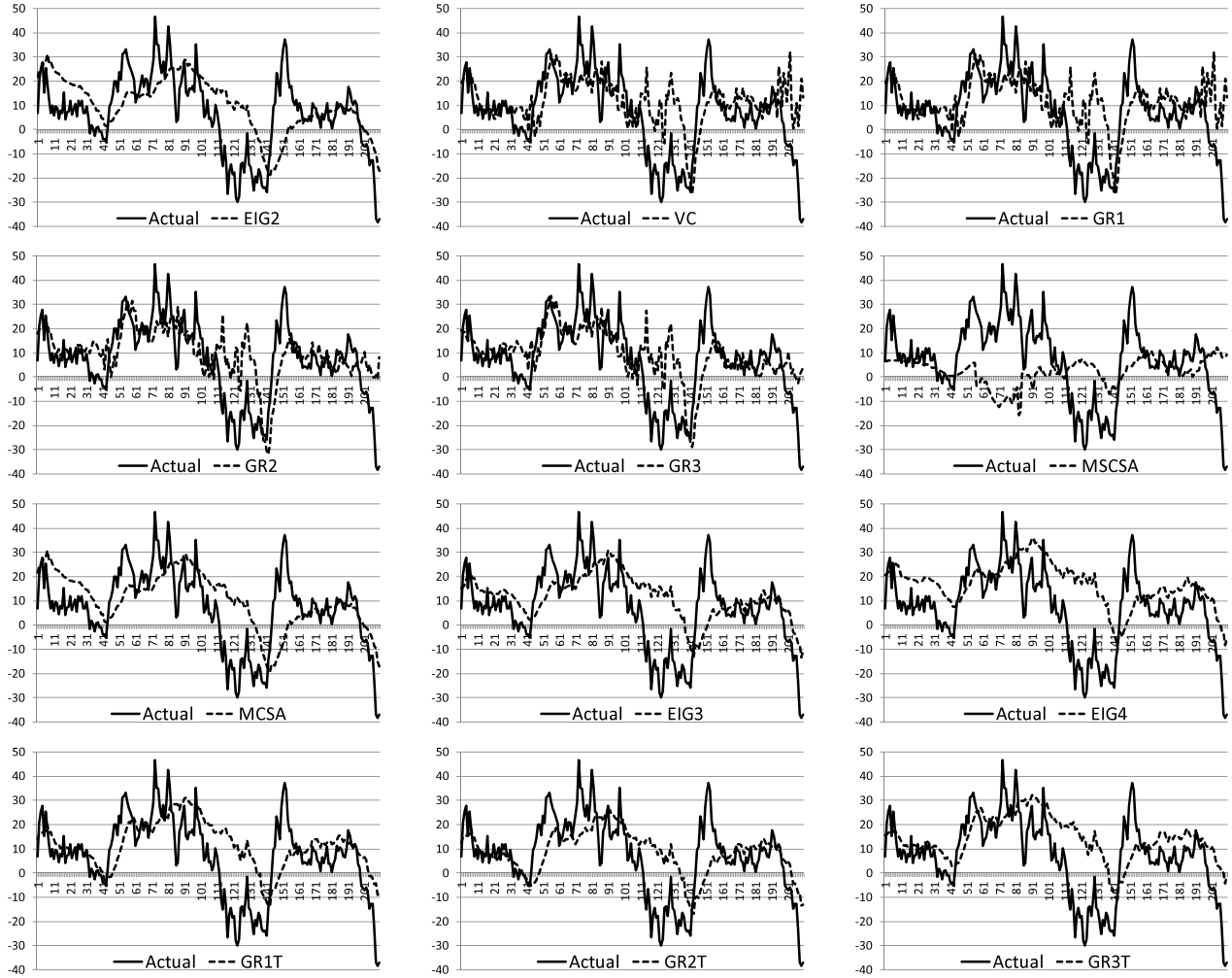

Fig. 5. (*continued*)

likely to be misspecified and may at best provide a reasonable "local" forecasts. If predictive models only approximate reality "locally", then it is conceivable that there will be frequent "breaks" from the modeling perspective, even the underlying structure is stable. In a world where there could be frequent breaks, the issue of optimally combining forecasts is an ill-posed question. It appears that a more relevant question should be the robustness of a combination method.

Both the Monte Carlo results and the applied examples appear to suggest that in the absence of reliable information about the break points and the size of the breaks, rolling window Bayesian averaging or a mean corrected simple average of all predictions is a robust way to deal with the chance events. Here, the window size is arbitrarily chosen. It is reasonable to expect that the performance of any forecasting combination methods will depend on the window size. Pesaran and Pick (forthcoming) have provided some useful guides for a random walk with a jump drift. It would be interesting to see if their approach can be generalized to more general cases.

References

Aiolfi, M., Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. J. Econometrics 135, 31–53.

Akaike, H., 1974. A new look at the statistical predictor identification. IEEE Trans. Automat. Control AC19, 716–723.

Baik, J., Silverstein, J.W., 2006. Eigenvalues of large sample covariance matrices of spiked population models. J. Multivariate Anal. 97, 1382–1408.

Bates, J.M., Granger, C.M.W., 1969. The combination of forecasts. Oper. Res. Q. 20, 451–468.

Bryan, M.F., Molloy, L., 2007. Mirror, mirror, who's the best forecaster of them all? In: Cleveland FRB Report.

Buckland, S.T., Burnhamn, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Capistran, C., Timmermann, A., 2009. Forecast combination with entry and exit of experts. J. Bus. Econom. Statist. 27, 428–440.

Chan, Y.L., Stock, J., Watson, M., 1999. A dynamic factor model framework for forecast combination. Spanish Econom. Rev. 1, 91–121.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. Int. J. Forecast. 5, 559–581.

Diebold, F.X., Pauly, P., 1990. The use of prior information in forecast combination. Int. J. Forecast. 6, 503–508.

Golub, G.H., Van Loan, C., 1980. An analysis of the total least squares problem. SIAM J. Numer. Anal. 17, 883–893.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financial Studies 21, 1455–1508.

Granger, C.W.J., Jeon, Y., 2004. Thick modeling. Econom. Model. 21, 323–343.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecast accuracy. J. Forecast. 19, 197–204.

Hansen, B.E., 2008. Least-squares forecast averaging. J. Econometrics 146, 342–350.

Hsiao, C., 2012. The creative tension between statistics and economics. Singapore Economic Review 57, 1–11.

Huang, H.Y., Lee, T.H., 2010. To combine forecasts or to combine information? Econom. Rev. 29, 534–570.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Markowitz, H., 1952. Portfolio selection. J. Finance 7, 77–99.

Markowitz, H., 1959. Portfolio Selection: Efficient Diversification of Investments. John Wiley, New York.

Nadler, B., 2008. Finite sample approximation results for principal component analysis: a matrix perturbation approach. Ann. Statist. 36, 2791–2817.

Newbold, P., Granger, C.W.J., 1974. Experience with forecasting univariate time series and the combination of forecasts. J. Roy. Statist. Soc. Ser. A 137, 131–165.
Palm, F.C., Zellner, A., 1992. To combine or not to combine? Issues of combining forecasts. J. Forecast. 11, 687–701.
Pesaran, M.H., Pick, A., 2010. Forecast combination across estimation windows. J. Bus. Econom. Statist. (forthcoming).
Schwartz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.
Stock, J.H., Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set. J. Forecast. 23, 405–430.
Swanson, N.R., Zeng, T., 2001. Choosing among competing econometric forecasts: regression-based forecast combination using model selection. J. Forecast. 20, 425–440.
Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Vol. 1. Elsevier, Amsterdam.