Journal of Econometrics 237 (2023) 105295

Contents lists available at ScienceDirect

Journal of Econometrics

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

# Optimal model averaging based on forward-validation

Xiaomeng Zhang [a,b], Xinyu Zhang [a,c,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Beijing Academy of Artificial Intelligence, Beijing 100084, China

ARTICLE INFO

*Article history:*
Received 20 November 2020
Received in revised form 6 January 2022
Accepted 20 March 2022
Available online 21 May 2022

*JEL classification:*
C22
C53

*Keywords:*
Model averaging
Forward-validation
Asymptotic optimality
Forecasting
Minimum risk
Window size

ABSTRACT

In this paper, noting that the prediction of time series follows the temporal order of data, we propose a frequentist model averaging method based on forward-validation. Our method also considers the uncertainty of the window size in estimation, i.e., we allow the sample size to vary among candidate models. We establish the asymptotic optimality of our method in the sense of achieving the lowest possible squared prediction risk. We also prove that if there exists one or more correctly specified models, our method will automatically assign all the weights to them. The promising performance of our method for finite samples is demonstrated by simulations and an empirical example of predicting the equity premium.

© 2022 Elsevier B.V. All rights reserved.

## 1. Introduction

As many multimodel methods have shown high prediction accuracy in empirical research, they have received increasing attention in recent years. Model averaging is one such method that is widely used in economic, financial, biological, medical and other fields. As discussed in Bates and Granger (1969) and Leung and Barron (2006), an average estimator often reduces the mean squared error in estimation because it avoids ignoring useful information about the form of the relationship between the response and covariates and provides a kind of insurance against selecting a very poor candidate model.

Model averaging methods can be categorized as Bayesian model averaging (BMA) or frequentist model averaging (FMA). For a comprehensive review of the BMA literature, see Hoeting et al. (1999). During the past two decades, FMA has received increasing attention and has experienced substantial development from both theoretical and empirical perspectives. FMA methods include averaging weights based on information criteria (see Buckland et al., 1997; Hjort and Claeskens, 2003, 2006; Zhang and Liang, 2011), optimal weighting (see Hansen, 2007; Wan et al., 2010; Liu and Okui, 2013; Lu and Su, 2015; Zhang, 2021), adaptive weighting (see Yang, 2001; Yuan and Yang, 2005; Zhang et al., 2013), the plug-in method (Liu, 2015; Lohmeyer et al., 2019), and others. Among them, the optimal weighting method is based on minimizing a weight-choosing criterion and has been proven to offer the minimal prediction loss in a large sample sense. There are several optimal weighting methods. Hansen and Racine (2012) proposed jackknife model averaging (JMA) under heteroscedasticity, where the weight-choosing criterion is built on the idea of the jackknife method, also

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
*E-mail address:* xinyu@amss.ac.cn (X. Zhang).

https://doi.org/10.1016/j.jeconom.2022.03.010
0304-4076/© 2022 Elsevier B.V. All rights reserved.

known as leave-one-out cross-validation. The leave-one-out cross-validation was also used in model averaging for Cox regression (He et al., 2020), expectile regression (Tu and Wang, 2020), vector autoregressive models (Liao and Tsay, 2020) and stochastic frontier models (Isaksson et al., 2021). Gao et al. (2016) proposed leave-subject-out model averaging (LsoMA), extending JMA by using leave-subject-out cross-validation in the weight-choosing procedure, and applied LsoMA to the prediction of time series. See Cheng and Hansen (2015), Cheng et al. (2015), Liu et al. (2020) and Hao et al. (Unpublished results) for more works on FMA methods which focus on the forecasting of various types of data and model structures. The idea of FMA is also applied to the generalized method of moments (GMM) estimation (Cheng et al., 2019), estimating the dynamic covariance matrix (Chen et al., 2019), analyzing clustered survival data (Xu et al., 2020), analyzing categorical data (Heiler and Mareckova, 2021) and many other cases. In addition, weighted average least squares (WALS) estimation, which combines the ideas of BMA and FMA, involves prior distributions and an analysis of estimation risk from a frequentist perspective; see Magnus et al. (2010, 2011) and De Luca et al. (2018). More reviews about model averaging and its application in economics can be seen in Ullah and Wang (2013) and Steel (2020).

In this article, we propose a new FMA method for forecasting related to cross-validation. We note that cross-validation is a commonly used weight-choosing technique in FMA; see the aforementioned papers. As indicated by Hjorth (1982), cross-validation makes efficient use of all the available data and requires no exact description of the distribution of the variables, only prediction formulas; however, there is an important condition that the predictions of excluded data in the sample should be made under certain assumptions which are approximately equivalent to those for the predictions of future observations made from the whole set of data. This condition can usually be fulfilled for independent variables but rules out many situations with strongly dependent data; see also Hjorth (1982). This characteristic may explain the possible invalidity of cross-validation for time series. Moreover, time series often have a natural temporal order, which is broken by cross-validation. Recently, Falessi et al. (2020) also emphasized the necessity of taking this temporal order into consideration when choosing validation techniques.

Forward-validation is a time series validation technique that can preserve the temporal order of time series forecasting, is widely used in data splitting and model evaluation in empirical studies and is also known as walk-forward validation in machine learning; see Kaastra and Boyd (1996). The basic idea for time series splitting is to divide the training set into two folds at each iteration on the condition that the validation set is always ahead of the training set. Consequently, this validation technique preserves the temporal order of the data, in contrast to cross-validation. To the best of our knowledge, theoretical work on forward-validation is not as rich as that on its empirical applications. One of the earliest works in statistics is Hjorth (1982). In time series literature, the idea of forward-validation is not new in out-of-sample performance evaluation; see Tashman (2000). As recommended by a referee, the forward-validation in Hjorth (1982) is essentially the same as the forecast evaluation with the recursive scheme (West, 2006). Welch and Goyal (2008) implemented their out-of-sample (OOS) evaluation on time series, which is very similar to forward-validation. In Welch and Goyal (2008), it states "The OOS forecast uses only the data available up to the time at which the forecast is made". Specifically, Welch and Goyal (2008) used the vectors of rolling OOS errors from the historical mean and the ordinary least squares (OLS) estimations to construct the OOS statistics for the OOS test and the model selection criterion. Different from Welch and Goyal (2008), we use the vectors of recursive OOS errors from the OLS estimations to build the weight choosing criterion. As far as we can know, our work is the first one that uses forward out-of-sample performance evaluation for the weight choice in model averaging. There are also some comparative studies on several forms of cross-validation and forward-validation for time series forecasting; see, for instance, Cerqueira et al. (2017) and Schnaubelt (2019).

In this article, we apply forward-validation to the model averaging method and propose a new optimal weighting model averaging method that is more in line with the characteristics of time series forecasting. We also provide two theoretical properties. First, the weight of our method minimizes the squared prediction risk asymptotically when all candidate models are misspecified. Second, if any candidate models are correctly specified, our procedure will automatically select them and assign no weight to the misspecified models.

Furthermore, we extend our method to allow for different window sizes in the candidate models. The window size here refers to the sample size in the estimation. Works on combining different window sizes can be found in Pesaran and Timmermann (2007), Pesaran and Pick (2011) and Jungmittag (2016), which focus on models with structural breaks and add equal weights to average the results obtained using different window sizes in the same model. In this article, we provide guidance to combine predictions from different models with different window sizes.

The remainder of this article is organized as follows. Section 2 presents the problem of interest and the model framework. Section 3 describes the procedure for obtaining the averaging weight based on forward-validation. Sections 4–5 establish the asymptotic properties of our procedure. The finite sample performance is investigated via numerical simulations in Section 6 and an empirical application of predicting the equity premium in Section 7. Section 8 concludes the article with some discussion. All technical details including detailed proofs are provided in Appendix A and the supplementary materials.

## 2. Model framework

Our approach begins by considering the data generating process (DGP)

$$y_{t+h} = \sum_{j=0}^{\infty} \theta_j y_{t-j} + \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \gamma_{j,k} x_{t-j,k} + e_{t+h}, \qquad t = 1, \ldots, T. \tag{1}$$

Considering the $h$-step-ahead forecast of $y_{t+h}$, the predictors are $y_t, x_{t,1}, \ldots, x_{t,\infty}$ and their lag terms. The error terms $\{e_{t+h}\}$ are independent unobserved variables satisfying

$$\mathbb{E}(e_{t+h} \mid y_t, x_{t,1}, \ldots, x_{t,\infty}, y_{t-1}, x_{t-1,1}, \ldots, x_{t-1,\infty}, \ldots) = 0.$$

Let $\sigma^2_{t+h} = \mathrm{var}(e_{t+h})$ denote the variance.

Since the DGP in Eq. (1) is unknown, econometricians often approximate it with a set of candidate models. Let $M$ be the number of candidate models; we treat $M$ as fixed in this article and the divergent case will be investigated in our future work. Notably, the DGP in Eq. (1) allows the existence of structural change. We give the following simple example to illustrate it. Suppose the DGP with a structural change is as follows,

$$y_{t+h} = \begin{cases} \theta_0 y_t + \theta_1 y_{t-1} + \gamma_{0,1} x_{t,1} + e_{t+h}, & t \le T_b; \\ \phi_0 y_t + \theta_1 y_{t-1} + \gamma_{0,1} x_{t,1} + e_{t+h}, & t > T_b. \end{cases} \tag{2}$$

Here, $T_b$ is a break point. Denote the index function $\mathbb{I}(t > T_b)$ as

$$\mathbb{I}(t > T_b) = \begin{cases} 0, & t \le T_b; \\ 1, & t > T_b. \end{cases}$$

Let $\gamma_{0,2} = \phi_0 - \theta_0$ and $x_{t,2} = y_t \mathbb{I}(t > T_b)$. Then, the DGP in Eq. (2) is equivalent to

$$y_{t+h} = \theta_0 y_t + \theta_1 y_{t-1} + \gamma_{0,1} x_{t,1} + \gamma_{0,2} x_{t,2} + e_{t+h}.$$

Therefore, our DGP can take the structural change into consideration. Generally, when predicting time series, it is not easy to determine the existence of the structural break. If the structure break exists, it is still uncertain whether there exists a single or multiple structural breaks, let alone the location of the break point, and using only a part of the full sample may attain better estimation results than using the full sample. Observations in different periods may carry different kinds of information, which plays an important role in time series analysis. Therefore, the uncertainty of the window size should not be neglected. Instead of choosing a single window size, we deal with the uncertainty of window size via averaging the estimators from candidate models with various window sizes. Specifically, in the $m$th candidate model, we use the most recent $T_m$ observations:

$$y_{t+h} = \sum_{j \in \widetilde{\mathcal{J}}_m} \theta_j^{(m)} y_{t-j} + \sum_{k \in \mathcal{K}_m} \sum_{j \in \mathcal{J}_m} \gamma_{j,k}^{(m)} x_{t-j,k} + e_{t+h}^{(m)}, \qquad t = T - T_m + 1, \ldots, T, \tag{3}$$

where, for $m \in \{1, \ldots, M\}$, $\mathcal{K}_m$ is the index set of covariates, $\widetilde{\mathcal{J}}_m$ and $\mathcal{J}_m$ are the sets including the order of the lags used in model $m$, $\theta_j^{(m)}$ and $\gamma_{j,k}^{(m)}$ denote the coefficients, and $e_{t+h}^{(m)}$ denotes the error term of model m. We need $T_m \ge h + |\widetilde{\mathcal{J}}_m| + |\mathcal{J}_m|$ with $|\widetilde{\mathcal{J}}_m|$ and $|\mathcal{J}_m|$, indicating the cardinal numbers of $\widetilde{\mathcal{J}}_m$ and $\mathcal{J}_m$, respectively. This restriction on $T_m$, which is quite common and mild, guarantees that the parameters are estimable. We allow the first term of the candidate model to be the intercept term. For model $m$, Eq. (3) can be written in matrix notation as

$$\mathbf{y}^{(m)} = \mathbf{Y}^{(m)} \boldsymbol{\beta}^{(m)} + \mathbf{e}^{(m)}, \tag{4}$$

where $\mathbf{y}^{(m)} = (y_{T-T_m+1}, \ldots, y_T)^\top$, $\mathbf{e}^{(m)} = (e_{T-T_m+1}^{(m)}, \ldots, e_T^{(m)})^\top$, $\mathbf{Y}^{(m)}$ is the matrix consisting of the variables $\{y_{t-h-i}, x_{t-h-j,k} \mid i \in \widetilde{\mathcal{J}}_m, j \in \mathcal{J}_m, k \in \mathcal{K}_m, t \in \{T - T_m + 1, \ldots, T\}\}$ [1] and $\boldsymbol{\beta}^{(m)}$ is the corresponding coefficient vector.

Under model $m$, the OLS estimator [2] of $\boldsymbol{\beta}^{(m)}$ is

$$\widehat{\boldsymbol{\beta}}^{(m)} = (\mathbf{Y}^{(m)\top} \mathbf{Y}^{(m)})^{-} \mathbf{Y}^{(m)\top} \mathbf{y}^{(m)}; \tag{5}$$

then, the $h$-step-ahead prediction of $y_{T+h}$ is

$$\widehat{y}_{T+h}^{(m)} = \mathbf{Y}_{T+h}^{(m)\top} \widehat{\boldsymbol{\beta}}^{(m)}, \tag{6}$$

where $\mathbf{Y}_{T+h}^{(m)}$ is a vector consisting of the variables $\{y_{T-i}, x_{T-j,k} \mid i \in \widetilde{\mathcal{J}}_m, j \in \mathcal{J}_m, k \in \mathcal{K}_m\}$ corresponding to coefficient vector $\boldsymbol{\beta}^{(m)}$.

Let weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)^\top$ be in the set

$$\mathcal{H} \equiv \left\{ \mathbf{w} = (w_1, w_2, \ldots, w_M)^\top \in [0, 1]^M : \sum_{m=1}^{M} w_m = 1 \right\}. \tag{7}$$

[1] We assume there are proper observations of $y_t, x_{t,k}$ before $t = T - T_m + 1$ such that the lengths of $\mathbf{y}^{(m)}$ and $\mathbf{Y}^{(m)}$ are all $T_m$.

[2] Although the DGP (1) allows heteroscedasticity, similar to Hansen and Racine (2012), we still utilize the OLS estimator because the variance matrix structure is unknown and the feasible generalized least squares (FGLS) estimator may not be better than OLS, but the calculation of FGLS is much more complicated than that of OLS.

The set $\mathcal{H}$ restricts all components of the weight vector to the interval [0, 1], and the sum of them to 1. The $h$-step-ahead model averaging prediction of $y_{T+h}$ has the following form

$$\widehat{y}_{T+h}(\mathbf{w}) = \sum_{m=1}^{M} w_m \widehat{y}_{T+h}^{(m)},$$

which is a weighted average of the predictions of individual candidate models.

## 3. Forward-Validation Model Averaging (FVMA) criterion

To choose the weight in $\widehat{y}_{T+h}(\mathbf{w})$, we propose a forward-validation criterion as follows.

Let $(\nu+1)$ be the earliest period used in the following forward-validation criterion (8). For any $m \in \{1, \ldots, M\}$ and $t \in \{\nu+1, \ldots, T\}$, we denote $\widetilde{y}_t^{(m)}$ as the $h$-step-ahead prediction of $y_t$ in model $m$ using only observations in the training set $\mathcal{T}_t^{(m)} = \{y_{t_0}, y_{t_0-h-i}, x_{t_0-h-j,k} \mid i \in \widetilde{\mathcal{J}}_m, j \in \mathcal{J}_m, k \in \mathcal{K}_m, t_0 \in \{T - T_m + 1, \ldots, t-1\}\}$. The calculation procedure is the same as that of the $h$-step-ahead prediction of $y_{T+h}$. Consequently, the model averaging prediction of $y_t$ combining these $M$ predictions is

$$\widetilde{y}_t(\mathbf{w}) = \sum_{m=1}^{M} w_m \widetilde{y}_t^{(m)}, \qquad (\nu+1) \le t \le T,$$

where $\nu > \max_{1\le m\le M}(T - T_m)$. Let $T_{min} = \min_{1\le m\le M} T_m$; then, there exist at least $\{T_{min} - (T-\nu)\}$ periods of data for every candidate model to estimate the coefficient vector, and $(T-\nu)$ periods for validation. To maintain a balance between the infimum of the sample size of the training set and the sample size of the validation set, we recommend choosing $\nu$ satisfying $\{T_{min} - (T-\nu)\}/(T-\nu) = c$, where c is a constant. We define the FVMA criterion as

$$\mathrm{FV}(\mathbf{w}) = \frac{1}{T-\nu} \sum_{t=\nu+1}^{T} \{y_t - \widetilde{y}_t(\mathbf{w})\}^2. \tag{8}$$

For candidate model $m \in \{1, \ldots, M\}$, let $\widetilde{e}_t^{(m)} = y_t - \widetilde{y}_t^{(m)}$, $\widetilde{\mathbf{e}}_{T-\nu}^{(m)} = (\widetilde{e}_{\nu+1}^{(m)}, \ldots, \widetilde{e}_T^{(m)})^\top$, matrix $\widetilde{\mathbf{e}}_{T-\nu} = (\widetilde{\mathbf{e}}_{T-\nu}^{(1)}, \ldots, \widetilde{\mathbf{e}}_{T-\nu}^{(M)})$ and $\mathbf{S}_{T-\nu} = \widetilde{\mathbf{e}}_{T-\nu}^\top \widetilde{\mathbf{e}}_{T-\nu}/(T-\nu)$; thus, we can write

$$\mathrm{FV}(\mathbf{w}) = \mathbf{w}^\top \mathbf{S}_{T-\nu} \mathbf{w}. \tag{9}$$

We use the FVMA criterion to select the weight vector $\widehat{\mathbf{w}}$ as follows:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}}{\arg\min}\ \mathrm{FV}(\mathbf{w}),$$

where $\mathbf{w}$ is restricted to the set $\mathcal{H}$ defined in Eq. (7). This optimization process based on Eq. (9) is a quadratic programming problem with respect to $\mathbf{w}$, so it can be solved easily and rapidly by various available software programs.

Compared with model averaging methods based on cross-validation, such as JMA and LsoMA, another advantage of our FVMA method is shown in the calculation of model averaging weights. When new observations are added to the training set, our FVMA method can update the model averaging weights much more simply than JMA and LsoMA because it can exploit the historical weight-choosing criterion, in contrast to model averaging methods based on cross-validation. When there is a new observation involved, each training set used in the weight-choosing procedure of model averaging methods based on cross-validation is updated; therefore, the new weight-choosing criterion cannot be built on the historical criterion. We adopt the following example to illustrate how the FVMA method updates the weight-choosing criterion. Suppose that we have already had observations until time period $T$ and calculated $\mathbf{S}_{T-\nu}$, and now a new observation at time period $T+1$ is available. Let $\widetilde{\mathbf{e}}_{new} = (e_{T+1}^{(1)}, \ldots, e_{T+1}^{(M)})$; then

$$\widetilde{\mathbf{e}}_{T+1-\nu} = \begin{pmatrix} \widetilde{\mathbf{e}}_{T-\nu} \\ \widetilde{\mathbf{e}}_{new} \end{pmatrix}.$$

Consequently,

$$\begin{aligned} \mathbf{S}_{T+1-\nu} &= \frac{1}{T+1-\nu} \widetilde{\mathbf{e}}_{T+1-\nu}^\top \widetilde{\mathbf{e}}_{T+1-\nu} \\ &= \frac{1}{T+1-\nu} \left(\widetilde{\mathbf{e}}_{T-\nu}^\top, \widetilde{\mathbf{e}}_{new}^\top\right) \begin{pmatrix} \widetilde{\mathbf{e}}_{T-\nu} \\ \widetilde{\mathbf{e}}_{new} \end{pmatrix} \\ &= \frac{1}{T+1-\nu} \left(\widetilde{\mathbf{e}}_{T-\nu}^\top \widetilde{\mathbf{e}}_{T-\nu} + \widetilde{\mathbf{e}}_{new}^\top \widetilde{\mathbf{e}}_{new}\right) \\ &= \frac{T-\nu}{T+1-\nu} \mathbf{S}_{T-\nu} + \frac{1}{T+1-\nu} \widetilde{\mathbf{e}}_{new}^\top \widetilde{\mathbf{e}}_{new}. \end{aligned} \tag{10}$$

From Eqs. (9) and (10), our FVMA method only needs to calculate $\widetilde{\mathbf{e}}_{new}$ additionally, which is the error of each method in predicting the new observation, to obtain the new value of the FVMA criterion. To describe our procedure more explicitly, we give the procedure as an algorithm, shown in Algorithm 1. We also give a simple example with $M = 2$ candidate models and specify what $\nu$ and $T_m$ are in this example in Section S.1 of the supplementary materials.

**Algorithm 1** Forward-validation model averaging (FVMA) algorithm

**Step 1. Calculate the $h$-step-ahead prediction of $y_{T+h}$ under every candidate model.**
For $m = 1$ to $M$:
Calculate $\widehat{\boldsymbol{\beta}}^{(m)}$ by Eq. (5); then calculate $\widehat{y}_{T+h}^{(m)}$ by Eq. (6).

**Step 2. Calculate the model averaging weight based on the forward-validation criterion.**

1. Compute $T_{min} = \min_{1\le m\le M} T_m$ and $\nu = T - T_{min}/(1+c)$.
2. For $t = \nu + 1$ to $T$:
   For $m = 1$ to $M$:
   (1) Split data into the training set and validation set.
   Training set: $\{y_{t_0}, y_{t_0-h-i}, x_{t_0-h-j,k} \mid i \in \widetilde{\mathcal{I}}_m, j \in \mathcal{J}_m, k \in \mathcal{K}_m, t_0 \in \{T - T_m + 1, \ldots, t-1\}\}$.
   Validation set: $\{y_t, y_{t-h-i}, x_{t-h-j,k} \mid i \in \widetilde{\mathcal{I}}_m, j \in \mathcal{J}_m, k \in \mathcal{K}_m\}$.
   (2) Using training set and following the method in **Step 1**, calculate the $h$-step-ahead prediction of $y_t$ under the $m$th candidate model, that is, $\widetilde{y}_t^{(m)}$.
   (3) Compute $\widetilde{e}_t^{(m)} = y_t - \widetilde{y}_t^{(m)}$.
3. Compute matrix $\widetilde{\mathbf{e}}_{T-\nu} \in \mathbb{R}^{(T-\nu)\times M}$ with $\widetilde{e}_{\nu+i}^{(j)}$ being the element in row $i$ column $j$, and $\mathbf{S}_{T-\nu} = \widetilde{\mathbf{e}}_{T-\nu}^{\top}\widetilde{\mathbf{e}}_{T-\nu}/(T-\nu)$.
4. Solve the constrained quadratic programming problem to obtain the model averaging weight
$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}}{\arg\min}\ \mathbf{w}^{\top}\mathbf{S}_{T-\nu}\mathbf{w}.$$

**Step 3. Calculate the $h$-step-ahead FVMA prediction of $y_{T+h}$.**
Let $\widehat{w}_m$ represent the $m$th element of $\widehat{\mathbf{w}}$; thus,
$$\widehat{y}_{T+h}(\widehat{\mathbf{w}}) = \sum_{m=1}^{M} \widehat{w}_m \widehat{y}_{T+h}^{(m)}.$$

## 4. Asymptotic optimality

In this section, we demonstrate that our FVMA method is asymptotically optimal in the sense that the prediction achieves the lowest possible risk when the sample size goes to infinity. To evaluate the performance of the FVMA method, we consider the risk function

$$R_{T+h}(\mathbf{w}) = \mathbb{E}\{y_{T+h} - \widehat{y}_{T+h}(\mathbf{w})\}^2 - \sigma_{T+h}^2. \tag{11}$$

As explained in Hansen (2008), the error variance $\sigma_{T+h}^2$ is subtracted because it is often the leading term in the mean-squared forecast error and is common across forecast methods. To illustrate this point in Hansen (2008), we decompose the mean-squared forecast error of prediction by OLS. Assume $\mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{e}$; then, under some very commonly used conditions, we have

$$\begin{aligned}
\mathbb{E}\{y_{T+h} - \widehat{y}_{T+h}\}^2 &= \mathbb{E}\left\{e_{T+h} + \mathbf{Y}_{T+h}^{\top}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\}^2 \\
&= \sigma_{T+h}^2 + \mathbb{E}\left\{\mathbf{Y}_{T+h}^{\top}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\}^2 \\
&= \sigma_{T+h}^2 + \frac{1}{T}\mathbb{E}\left\{\mathbf{Y}_{T+h}^{\top}\left(\frac{1}{T}\mathbf{Y}^{\top}\mathbf{Y}\right)^{-1}\left(\frac{1}{\sqrt{T}}\mathbf{Y}^{\top}\mathbf{e}\right)\right\}^2 \\
&= \sigma_{T+h}^2 + o(1).
\end{aligned}$$

To present the asymptotic optimality, we need some regularity conditions. Unless otherwise stated, all limiting properties are set in $(T-\nu)\to\infty$.

**Assumption 1.** For any $m\in\{1,\ldots,M\}$, there exists a vector $\boldsymbol{\beta}^{*(m)}$ such that $\widehat{\boldsymbol{\beta}}^{(m)}-\boldsymbol{\beta}^{*(m)}=O_p\left((T-\nu)^{-1/2}\right)$.

Clearly, when the candidate model is correctly specified, $\boldsymbol{\beta}^{*(m)}$ is the true coefficient vector. For the misspecified candidate model, White (1982) proves the existence of $\boldsymbol{\beta}^{*(m)}$ and convergence rate $(T-\nu)^{-1/2}$ under regularity conditions. Let model $m_0$ be the index of the candidate model with the minimum sample size, i.e., $T_{min}=T_{m_0}$. When calculating the estimator at the earliest time period under the $m_0$th model, i.e., $\widetilde{y}^{(m_0)}_{\nu+1}$, our method uses the minimum sample size $\{T_{min}-(T-\nu)\}$. As we recommend choosing $\nu$ satisfying $\{T_{min}-(T-\nu)\}/(T-\nu)=c$ in the paragraph above Eq. (8), all the estimations of parameter $\boldsymbol{\beta}$ in the FV($\mathbf{w}$) share the same convergence rate in Assumption 1; see more detailed analyses in the antepenultimate paragraph of Appendix A. Here, we treat the dimension of $\boldsymbol{\beta}^{*(m)}$ as fixed, and the divergent case will be analyzed in our future study.

**Assumption 2.** Uniformly for $t\in\{1,\ldots,T+h\}$, $k\in\{1,\ldots,\infty\}$ and $m\in\{1,\ldots,M\}$, $y_t$ and $x_{t,k}$ are all $O_p(1)$ and $\boldsymbol{\beta}^{*(m)}$ is $O(1)$.

Assumption 2 is on the boundedness of observations and limiting parameters, and is a mild technical condition.

For $t\in\{\nu+1,\ldots,T+h\}$, let $y^{*(m)}_t=\mathbf{Y}^{(m)\top}_t\boldsymbol{\beta}^{*(m)}$, where $\mathbf{Y}^{(m)}_t$ is a vector consisting of the variables $\{y_{t-h-i},x_{t-h-j,k}|\ i\in\widetilde{\mathcal{J}}_m,j\in\mathcal{J}_m,k\in\mathcal{K}_m\}$ corresponding to coefficient vector $\boldsymbol{\beta}^{(m)}$, $e^{*(m)}_t=y_t-y^{*(m)}_t$, $y^*_t(\mathbf{w})=\sum_{m=1}^M w_m y^{*(m)}_t$, and the corresponding "risk" $R^*_{T+h}(\mathbf{w})=\mathbb{E}\left\{y_{T+h}-y^*_{T+h}(\mathbf{w})\right\}^2-\sigma^2_{T+h}$. Let $\xi^*_{T+h}=\inf_{\mathbf{w}\in\mathcal{H}}R^*_{T+h}(\mathbf{w})$ be the minimum risk under limiting parameter values. We further need the following conditions.

**Assumption 3.**

(i) $\sup_T\xi^*_{T+h}<\infty$;
(ii) $\xi^{*-1}_{T+h}(T-\nu)^{-1/2}=o(1)$;
(iii) $\sup_T\mathbb{E}\left\{\xi^{*-1}_{T+h}\max_{1\le i\le M}\left|\widehat{y}^{(i)}_{T+h}-y^{*(i)}_{T+h}\right|\right\}^{2+\delta_1}<\infty$ for some constant $\delta_1>0$.

Assumption 3 is on the minimum risk under limiting parameter values. Assumption 3(i) requires that the potential best prediction always has finite risk when the sample size varies. Assumption 3(ii) requires all the candidate models to be misspecified because if the $j$th candidate model is correctly specified, then $\widehat{\boldsymbol{\beta}}^{(j)}$ will converge to the true coefficient, so we have

$$\xi^*_{T+h}=\inf_{\mathbf{w}\in\mathcal{H}}R^*_{T+h}(\mathbf{w})\le R^*_{T+h}(\mathbf{w}^{(j)}_0)=\mathbb{E}\left\{y_{T+h}-y^{*(j)}_{T+h}\right\}^2-\sigma^2_{T+h}=0, \tag{12}$$

where $\mathbf{w}^{(j)}_0$ is a vector whose $j$th component is 1 and all other components are 0. Clearly, conclusion (12) conflicts with Assumption 3(ii). Since the data generating process is usually complex and unknown in real life, it is difficult to ascertain which variables are useful for prediction. Furthermore, it is also difficult to determine the functional form in which the response variable depends on the predictors. This determination process may result in problems, such as omitting variables, specifying a wrong functional form and so on. Therefore, the model misspecification required by Assumption 3(ii) can be satisfied in reality and is mild. Although the requirement of misspecification can be removed if we use $R_{T+h}(\mathbf{w})+\sigma^2_{T+h}$ to evaluate the prediction performance of the method, instead of the risk function $R_{T+h}(\mathbf{w})$ in Eq. (11), we recommend not analyzing the asymptotic optimality in Theorem 1 based on $R_{T+h}(\mathbf{w})+\sigma^2_{T+h}$ because it will greatly reduce the significance of asymptotic optimality. We state the asymptotic optimality based on $R_{T+h}(\mathbf{w})+\sigma^2_{T+h}$ in Theorem S1 and give its proof in Section S.2 of the supplemental materials, where we also provide more discussions on this issue. In the next section, we illustrate that when there exists one or more correctly specified candidate models, our method assigns all the weights to correctly specified models. Assumption 3(iii) is a useful sufficient technical condition for uniform integrability. In the proof of Theorem 1, we can prove $\xi^{*-1}_{T+h}\max_{1\le i\le M}\left|\widehat{y}^{(i)}_{T+h}-y^{*(i)}_{T+h}\right|=o_p(1)$ under other assumptions; thus, Assumption 3(iii) is actually mild.

**Assumption 4.** For any $m\in\{1,\ldots,M\}$, $\{y_t,\mathbf{Y}^{(m)}_t\}$ is a mixing sequence with either $\phi$ of size $-r/(2r-1)$, $r\ge 2$ or $\alpha$ of size $-r/(r-2)$, $r>2$.

Assumption 4 describes the dependence of the data. The assumption is similar to Condition 8 (i) in Gao et al. (2016), weaker than Assumption 3.2 in Liu and Kuo (2016) and much weaker than the i.i.d. requirement of Theorem 2 in Hansen (2008).

From the definition of $e^{*(m)}_t$, it can be interpreted as the bias of model $m$ for predicting $y_t$ when $(T-\nu)$ goes to infinity. Since Assumption 3(ii) implies that all the candidate models are misspecified, according to Eqs. (1) and (3), we know that $e^{*(m)}_t$ is a linear combination of some variables in $\{y_{t-h},x_{t-h,1},\ldots,x_{t-h,\infty},y_{t-h-1},x_{t-h-1,1},\ldots,x_{t-h-1,\infty},\ldots\}$ plus the error term $e_t$. The required conditions on these bias terms are as follows.

**Assumption 5.** There exist positive constants $c_1$, $c_2$ and $\delta_2$ such that

(i) $\mathrm{var}\left((T-\nu)^{-1/2}\sum_{t=\nu+1}^{T} e_t^{*(i)}e_t^{*(j)}\right) \geq c_1 > 0$ for all $(T-\nu)$ sufficiently large, for any $i,j \in \{1,\ldots,M\}$;

(ii) $\mathrm{var}\left\{ e_t^{*(i)}e_t^{*(j)}\right\} \leq c_2 < \infty$, $(\nu+1) \leq t \leq T$ for any $i,j \in \{1,\ldots,M\}$;

(iii) $\sup_T \mathbb{E}\left\{\xi_{T+h}^{*-1}\max_{1\leq i,j\leq M}\left|\widehat{y}_{T+h}^{(i)} - y_{T+h}^{*(i)}\right|\left|e_{T+h}^{*(j)}\right|\right\}^{2+\delta_2} < \infty$;

(iv) $\left|(T-\nu)^{-1}\sum_{t=\nu+1}^{T}\mathbb{E}e_t^{*(i)}e_t^{*(j)} - \mathbb{E}e_{T+h}^{*(i)}e_{T+h}^{*(j)}\right| = O((T-\nu)^{-1/2})$ for any $i,j \in \{1,\ldots,M\}$.

Assumption 5(i)–(ii) are regularity conditions of the central limit theorem for dependent processes; see Schönfeld (1971), Scott (1973), McLeish (1974), Heyde (1975) and Wooldridge and White (1988). Assumption 5(i) ensures that the variance does not degenerate to zero as $(T-\nu)$ goes to infinity. Assumption 5(ii) restricts the bound of the moment. Assumption 5(iii) is to guarantee the integrability of $\xi_{T+h}^{*-1}\max_{1\leq i,j\leq M}\left|\widehat{y}_{T+h}^{(i)} - y_{T+h}^{*(i)}\right|\left|e_{T+h}^{*(j)}\right|$ via moments. From our assumptions and proof steps, we have $\max_{1\leq i\leq M}\left|e_{T+h}^{*(i)}\right| = O_p(1)$ and $\xi_{T+h}^{*-1}\max_{1\leq i\leq M}\left|\widehat{y}_{T+h}^{(i)} - y_{T+h}^{*(i)}\right| = o_p(1)$, so $\xi_{T+h}^{*-1}\max_{1\leq i,j\leq M}\left|\widehat{y}_{T+h}^{(i)} - y_{T+h}^{*(i)}\right|\left|e_{T+h}^{*(j)}\right| = o_p(1)$. Thus, Assumption 5(iii) is a mild requirement. Assumption 5(iv) can be understood as a request for the rationality of using past information to predict $y_{T+h}$ and is much milder than the i.i.d. assumption that is required in Theorem 2 of Hansen (2008), because Assumption 5(iv) allows for stationary and some kind of nonstationary conditions. For example, when $\{y_t, x_{t,k}\}$ is weakly stationary, the first-order and second-order moments are correlated only with time interval, so the left side of the equation in Assumption 5(iv) is equal to zero. Furthermore, if all the candidate models contain the nonstationary part in Eq. (1) completely, leaving only the stationary variables as the components of $e_t^{*(m)}$, then the left side of the equation in Assumption 5(iv) will also equal zero. In particular, some kind of nonstationary components may also be allowed to exist in $e_t^{*(m)}$ as long as the time-averaged variance and covariance of the bias terms are close to the corresponding values at the time point $(T+h)$.

**Theorem 1.** *If Assumptions 1–5 hold, then*

$$\frac{R_{T+h}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} R_{T+h}(\mathbf{w})} \xrightarrow{p} 1.$$

This theorem demonstrates that the prediction made by our FVMA method is optimal in the sense that its squared risk is asymptotically identical to that of the infeasible best model averaging prediction. See Appendix A for the proof of Theorem 1.

## 5. Convergence of weights

In Section 4, we focus on the theoretical property of our method when all candidate models are misspecified. In contrast, we now consider the case where there exists one or more correctly specified candidate models. In this section, we demonstrate that under this case, without knowing which candidate models are correctly specified, our method can automatically assign all the weights to the predictions made by these correctly specified models.

Let $\mathcal{D}$ be the subset of $\{1,\ldots,M\}$ that consists of the indices of the correctly specified candidate models and $\mathcal{D}^c$ be the complement of $\mathcal{D}$. Let $\tau(\mathbf{w}) = \sum_{j\in\mathcal{D}} w_j$, $\widehat{w}_j$ be the $j$th element of $\widehat{\mathbf{w}}$, $\tau(\widehat{\mathbf{w}}) = \sum_{j\in\mathcal{D}}\widehat{w}_j$ and $\widetilde{\xi}_{T+h}^{*} = \inf_{\mathbf{w}\in\mathcal{H},\tau(\mathbf{w})=0} R_{T+h}^{*}(\mathbf{w})$. We further need the following assumption.

**Assumption 6.** $\widetilde{\xi}_{T+h}^{*-1}(T-\nu)^{-1/2} = o(1)$.

Clearly, when all the models are misspecified, $\mathcal{D}$ is an empty set; thus, $\widetilde{\xi}_{T+h}^{*} = \xi_{T+h}^{*}$. From this perspective, Assumption 3(ii) can be viewed as a special case of Assumption 6.

**Theorem 2.** *If there is one or more correctly specified estimation models and Assumption 1, 2, 4, 5(i), 5(ii), 5(iv) and 6 are satisfied, then*

$$\tau(\widehat{\mathbf{w}}) \xrightarrow{p} 1.$$

This theorem demonstrates that our FVMA method can select all the correctly specified models asymptotically, which is similar to the consistency of model selection. See Appendix B for the proof of Theorem 2.

## 6. Simulation study

Since Section 4 analyzes the asymptotic optimality in the case where all candidate models are misspecified and Section 5 explores the convergence of weights in the other case where one or more candidate models are correctly specified, in this simulation study, we consider two corresponding designs to evaluate the finite sample performance of our FVMA method in comparison with commonly used approaches.

**Table 1**
Candidate models in the simulation study.

| | Predictors |
|---|---|
| $m_0$ | 1 |
| $m_1$ | $y_t$ |
| $m_2$ | $y_t,\ x_t$ |
| $m_3$ | $y_t,\ x_t,\ x_{t-1}$ |
| $m_4$ | $y_t,\ x_t,\ x_{t-1},\ x_{t-2}$ |
| $m_5$ | $y_t,\ y_{t-1}$ |
| $m_6$ | $y_t,\ y_{t-1},\ x_t$ |
| $m_7$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}$ |
| $m_8$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}, x_{t-2}$ |
| $m_9$ | $y_t,\ y_{t-1},\ y_{t-2}$ |
| $m_{10}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t$ |
| $m_{11}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1}$ |
| $m_{12}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1},\ x_{t-2}$ |

### 6.1. Simulation design

Similar to the simulation design in Ng (2013) and Liu and Kuo (2016), we use a general DGP:

$$y_{t+h} = A(L)y_t + B(L)x_t + e_{t+h},$$

$$x_t = \rho_x x_{t-1} + u_t$$

and

$$\begin{pmatrix} e_t \\ u_{t'} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

where $L$ is lag operator, polynomial $A(z) = \phi_0 + \phi_1 z$ with $\phi_0 = 0.5$, $\phi_1$ is varied on a grid from $-0.5$ to 0.5, and polynomial $B(z) = \psi_0 + \psi_1 z$ with $\psi_0 = 0.8$, while $\psi_1$ takes different values under the two designs, which is the only difference between the two designs. The exogenous predictor $x_t$ is an AR(1) process with $\rho_x = 0.5$.

We use 13 linear models with all $T$ observations as candidate models and list the predictors in Table 1. Moreover, we focus on not only one-step-ahead prediction but also multi-step-ahead, setting $h \in \{1, 2, 4\}$ in each design.

**Design 1** (*All the Candidate Models are Misspecified*)**.** In this design, the DGP is subject to multiple structural breaks at time $T_{b1} = 80$, $T_{b2} = 160$, $T_{b3} = 240$ and $T_{b4} = 360$, which is not a rare structure in economics. Specifically, we generate the data following the setting in Section 6.1 and let the parameter

$$\psi_1 = \begin{cases} 0.7, & t \le T_{b1} \\ 0.2, & T_{b1} < t \le T_{b2} \\ -0.1, & T_{b2} < t \le T_{b3} \\ -0.4, & T_{b3} < t \le T_{b4} \\ -0.7, & t > T_{b4} \end{cases}.$$

We set the sample size $T \in \{100, 200, 400\}$. From the example in the second paragraph of Section 2, it can be seen that the setup in Design 1 is allowed by the DGP in Eq. (1). Since every candidate model omits the structural change, all the candidate models listed in Table 1 are misspecified.

**Design 2** (*There is One or More Correctly Specified Candidate Models*)**.** In this design, we generate the data in the same way as in Design 1 but let the parameter

$$\psi_1 = 0$$

at all times. We set the sample size $T \in \{100, 200, 400, 800, 1600\}$. Therefore, models $m_6$–$m_8$ and $m_{10}$–$m_{12}$ are always correctly specified, regardless of the value of $\phi_1$, while models $m_2$–$m_4$ are correctly specified only if $\phi_1 = 0$ and the others are misspecified all the time. Here, we consider larger sample sizes than those in Design 1 to check the convergence of the weights.

### 6.2. Simulation results

To evaluate the prediction performance of the method, we calculate the mean-squared forecast error (MSFE)

$$\mathrm{MSFE} = \frac{1}{D}\sum_{r=1}^{D}\{y_{t+h,r} - \widehat{y}_{t+h,r}\}^2, \tag{13}$$

**Table 2**
Candidate models for $\text{FVMA}_{49}$ in the simulation study.

| | Predictors | $T_m$ | | Predictors | $T_m$ |
|---|---|---|---|---|---|
| $m_0$ | 1 | | | | |
| $m_1$ | $y_t$ | | $m_{25}$ | $y_t$ | |
| $m_2$ | $y_t,\ x_t$ | | $m_{26}$ | $y_t,\ x_t$ | |
| $m_3$ | $y_t,\ x_t,\ x_{t-1}$ | | $m_{27}$ | $y_t,\ x_t,\ x_{t-1}$ | |
| $m_4$ | $y_t,\ x_t,\ x_{t-1},\ x_{t-2}$ | | $m_{28}$ | $y_t,\ x_t,\ x_{t-1},\ x_{t-2}$ | |
| $m_5$ | $y_t,\ y_{t-1}$ | $T$ | $m_{29}$ | $y_t,\ y_{t-1}$ | $0.6T$ |
| $m_6$ | $y_t,\ y_{t-1},\ x_t$ | | $m_{30}$ | $y_t,\ y_{t-1},\ x_t$ | |
| $m_7$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}$ | | $m_{31}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}$ | |
| $m_8$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}, x_{t-2}$ | | $m_{32}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}, x_{t-2}$ | |
| $m_9$ | $y_t,\ y_{t-1},\ y_{t-2}$ | | $m_{33}$ | $y_t,\ y_{t-1},\ y_{t-2}$ | |
| $m_{10}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t$ | | $m_{34}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t$ | |
| $m_{11}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1}$ | | $m_{35}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1}$ | |
| $m_{12}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1},\ x_{t-2}$ | | $m_{36}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1},\ x_{t-2}$ | |
| | Predictors | $T_m$ | | Predictors | $T_m$ |
| $m_{13}$ | $y_t$ | | $m_{37}$ | $y_t$ | |
| $m_{14}$ | $y_t,\ x_t$ | | $m_{38}$ | $y_t,\ x_t$ | |
| $m_{15}$ | $y_t,\ x_t,\ x_{t-1}$ | | $m_{39}$ | $y_t,\ x_t,\ x_{t-1}$ | |
| $m_{16}$ | $y_t,\ x_t,\ x_{t-1},\ x_{t-2}$ | | $m_{40}$ | $y_t,\ x_t,\ x_{t-1},\ x_{t-2}$ | |
| $m_{17}$ | $y_t,\ y_{t-1}$ | | $m_{41}$ | $y_t,\ y_{t-1}$ | |
| $m_{18}$ | $y_t,\ y_{t-1},\ x_t$ | $0.8T$ | $m_{42}$ | $y_t,\ y_{t-1},\ x_t$ | $0.4T$ |
| $m_{19}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}$ | | $m_{43}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}$ | |
| $m_{20}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}, x_{t-2}$ | | $m_{44}$ | $y_t,\ y_{t-1},\ x_t,\ x_{t-1}, x_{t-2}$ | |
| $m_{21}$ | $y_t,\ y_{t-1},\ y_{t-2}$ | | $m_{45}$ | $y_t,\ y_{t-1},\ y_{t-2}$ | |
| $m_{22}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t$ | | $m_{46}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t$ | |
| $m_{23}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1}$ | | $m_{47}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1}$ | |
| $m_{24}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1},\ x_{t-2}$ | | $m_{48}$ | $y_t,\ y_{t-1},\ y_{t-2},\ x_t,\ x_{t-1},\ x_{t-2}$ | |

where $D = 1000$ is the number of replications, $y_{t+h,r}$ is $y_{t+h}$ in the $r$th replication and $\widehat{y}_{t+h,r}$ is the prediction of $y_{t+h,r}$.

We compare the performance of our FVMA method with that of some commonly used approaches, including Mallows model averaging (MMA) proposed by Hansen (2007), JMA proposed by Hansen and Racine (2012), LsoMA proposed by Gao et al. (2016), multivariate leave-$h$-out cross-validation averaging ($\text{MCVA}_h$) proposed by Liao and Tsay (2020), the model with all the predictors (Full Model), lasso and random forest (RF). When implementing our method, we set the parameter $\nu = T - 0.8T_{min}$ in (8) in both Sections 6 and 7. We have also tried other settings for $\nu$ and obtained similar numerical results to those in this paper. The tuning parameters in lasso are selected by 10-fold cross-validation. We apply RF via the toolbox in MATLAB with the hyperparameters set to the default values. Let $\text{FVMA}_{13}$ denote our FVMA method combining these 13 candidate models in Table 1.

Fig. 1 depicts the MSFE values under Design 1. When all the candidate models are misspecified, our $\text{FVMA}_{13}$ method outperforms the other model averaging methods and Full Model under an overwhelming majority of configurations, except when $T = 400$, $h = 1$ and $\phi_1 = 0.5$. This superiority reflects the benefits of considering the temporal order of data in time series forecasting. $\text{FVMA}_{13}$ outperforms Lasso and RF methods under every setting.

Next, we consider multiple window sizes for our method. Specifically, we let $T_m$ vary among $\{T, 0.8T, 0.6T, 0.4T\}$ and extend the 13 ($=1 + 12$) candidate models in Table 1 to the 49 ($=1 + 12 \times 4$) candidate models in Table 2. Fig. 1 shows that $\text{FVMA}_{49}$ performs much better than all the other methods regardless of the values of $\phi_1$, $T$ and $h$. From the comparison between $\text{FVMA}_{13}$ and $\text{FVMA}_{49}$, this promotion in forecast accuracy can be attributed to the fact that $\text{FVMA}_{49}$ considers the uncertainty of the window size, while $\text{FVMA}_{13}$ does not. Additionally we run another three methods in our numerical examples: the forecast combination method using equal weight (Equal), adaptive boosting (Adaboost) and historical average (HA). These methods perform worse than our method in all configurations, so we do not present them for space considerations.

Fig. 2 depicts the MSFE values under Design 2, in which there is one or more correctly specified candidate models. RF shows obvious inferiority and lasso has the penultimate performance in all configurations. The other methods have similar performance, especially when $T$ is large. The performance of $\text{FVMA}_{13}$ is very similar to that of MMA, and even slightly worse than MMA sometimes in the setup of Design 2. Thus, to further compare these two methods, we add another simulation study following the setup in Hansen (2008), and the results show that $\text{FVMA}_{13}$ outperforms MMA under all parameter settings, although the difference is not obvious neither; see Section S.3 of the supplemental materials for details. When the sample size $T$ is not large, $\text{FVMA}_{49}$ performs worse than MMA and their performances become much closer as the sample size increases. As $\text{FVMA}_{49}$ uses 49 models while $\text{FVMA}_{13}$ and MMA use only 13 models (the correct models with the largest number of observations are in these 13 models), some disturbance from the other 36 candidate models may cause this disadvantage in MSFE shown in Fig. 2 when the sample size is not large. Noting that the asymptotic optimality in Theorem 1 requires that all the candidate models are misspecified (we think this situation is very common in practice), we cannot, theoretically, guarantee the asymptotic optimality when one or more correctly specified models

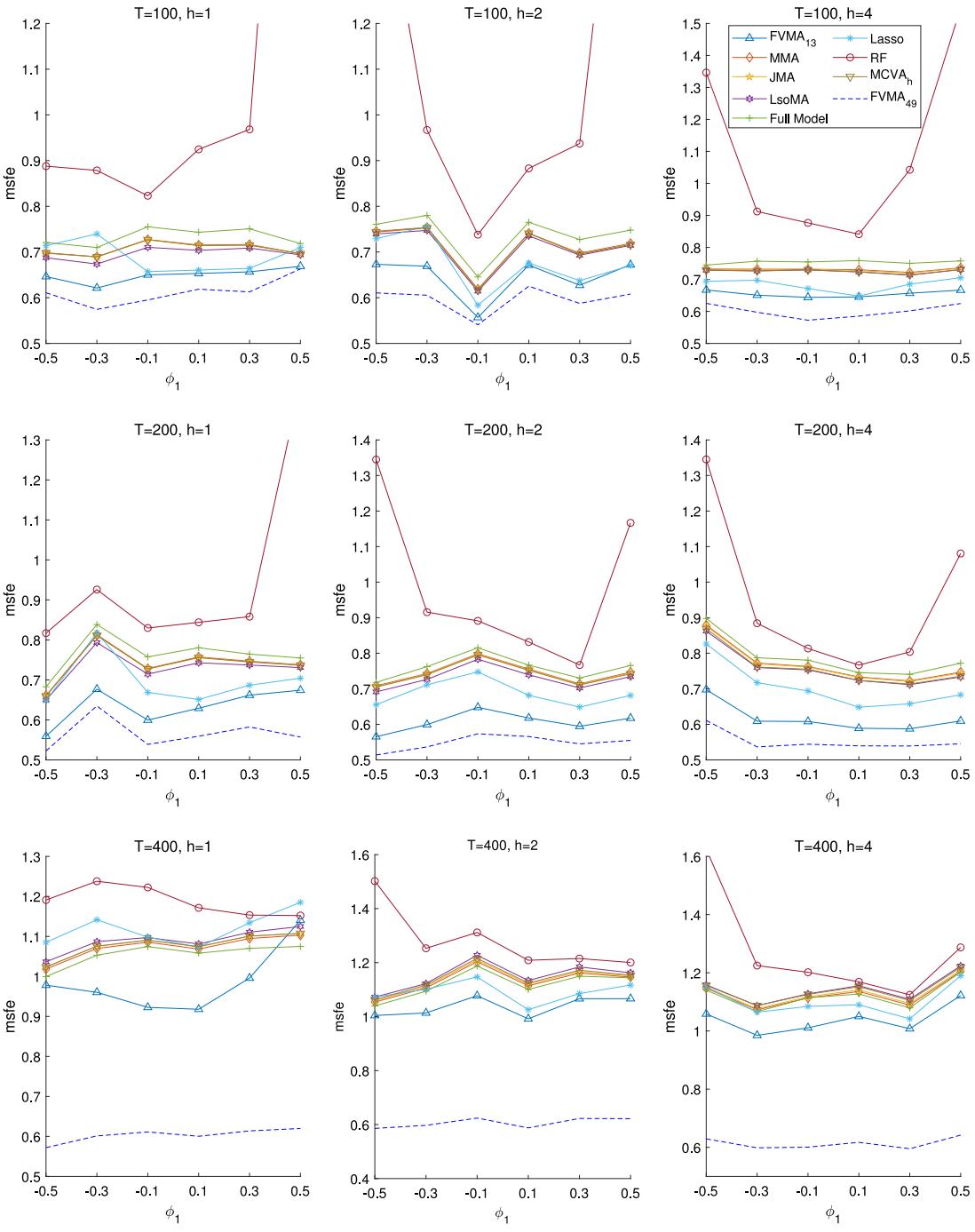


**Fig. 1.** MSFE values of 9 methods in Design 1.

exist. The asymptotic optimality when there exists one or more correctly specified models is a quite interesting and meaningful problem to investigate in future studies. In the current paper, we only focus on the consistency of the model averaging weight when there exists one or more correctly specified models, and the corresponding simulation result is shown in Fig. 3. Fig. 3 presents the sum of weights allocated to correctly specified candidate models by the FVMA method in Design 2. The sum converges to one as $T$ increases, which coincides with the finding in Theorem 2.

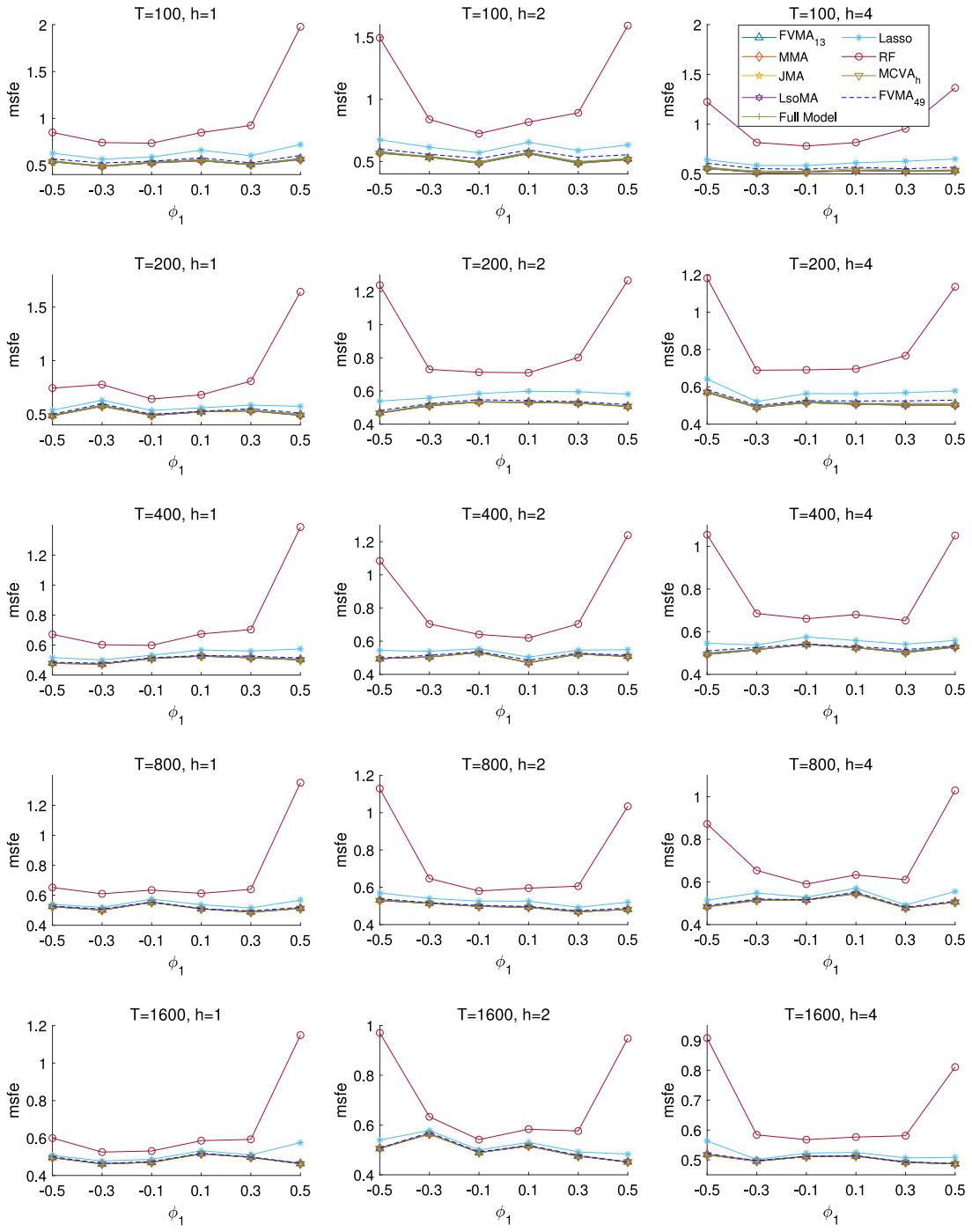


**Fig. 2.** MSFE values of 9 methods in Design 2.

## 7. Empirical application

In this section, we apply our method to predict the annual excess equity premium over the S&P 500 index and compare the prediction performance with that of all the methods mentioned in Section 6. Many studies have addressed equity premium prediction with various models and predictors, among which Welch and Goyal (2008) argue that numerous economic variables and models fail to achieve consistently good out-of-sample prediction performance relative to the HA method. However, Welch and Goyal (2008) ignore model uncertainty, which may cause the poor prediction

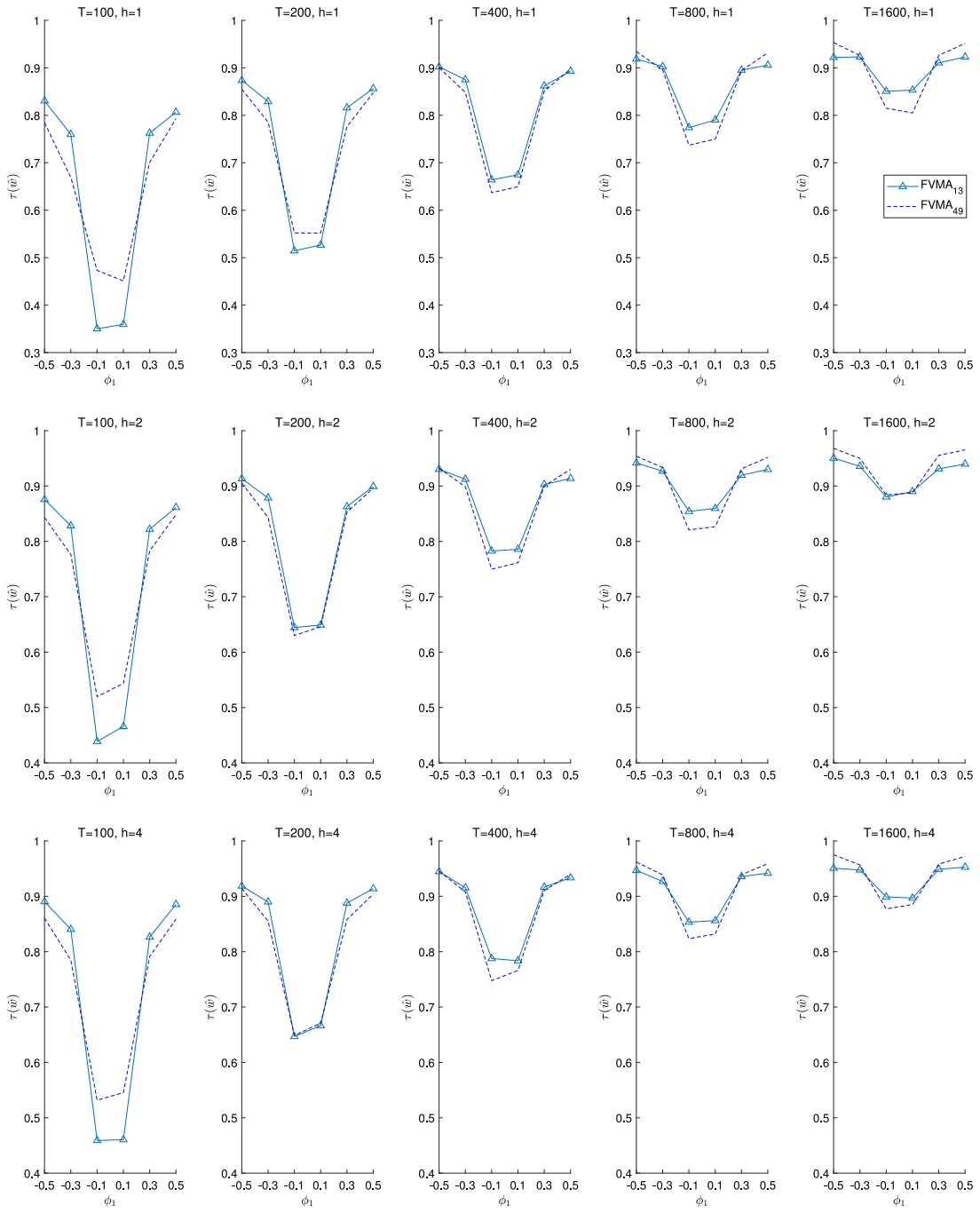


**Fig. 3.** Weight of FVMA allocated to correctly specified candidate models in Design 2.

performance. In contrast, Rapach et al. (2009), Liu and Kuo (2016), Wang et al. (2021) and many other works use forecast combination methods in equity premium prediction. Our method considers not only model uncertainty but also window size uncertainty.

### 7.1. Dataset and model setting

For $h$-step-ahead prediction of the equity premium, we use a linear model

$$EP_{t+h} = \mathbf{Z}_t^\top \boldsymbol{\beta} + e_{t+h},$$

**Table 3**
Average and standard error of the SFE.

| SFE ($\times 10^{-2}$) | $h=1$ | | $h=2$ | | $h=4$ | |
|---|---|---|---|---|---|---|
| | Average | s.e. | Average | s.e. | Average | s.e. |
| $\text{FVMA}_{31}$ | 0.3368② | 0.0315 | 0.2518② | 0.0752 | 0.5295③ | 0.1160 |
| MMA | 0.3397 | 0.0328 | 0.2558③ | 0.0752 | 0.5436 | 0.1150 |
| JMA | 0.3379 | 0.0322 | 0.2601 | 0.0750 | 0.5281② | 0.1125 |
| LsoMA | 0.3380 | 0.0319 | 0.2591 | 0.0747 | 0.5530 | 0.1149 |
| $\text{MCVA}_h$ | 0.3378③ | 0.0322 | 0.2595 | 0.0747 | 0.5434 | 0.1136 |
| $\text{FVMA}_{121}$ | 0.3183① | 0.0300 | 0.2103① | 0.0722 | 0.5104① | 0.1123 |
| Full Model | 0.3521 | 0.0350 | 0.2845 | 0.0772 | 0.5678 | 0.1227 |
| Lasso | 0.3710 | 0.0308 | 0.2906 | 0.0751 | 0.6510 | 0.1302 |
| RF | 0.6826 | 0.0519 | 0.6384 | 0.0904 | 0.9927 | 0.1326 |
| $\text{Equal}_{121}$ | 0.4816 | 0.0331 | 0.4218 | 0.0838 | 0.8176 | 0.1451 |
| AdaBoost | 1.7823 | 0.1690 | 2.3840 | 0.3062 | 2.6458 | 0.2706 |
| HA | 2.4466 | 0.1791 | 1.8957 | 0.1858 | 1.4919 | 0.1860 |

The superscripts ①, ② and ③ represent the best, second best and third best values in each column, respectively.

where $EP_{t+h}$ is the equity premium, $\mathbf{Z}_t$ contains predictors, and $e_{t+h}$ is an unobservable disturbance term. We set $h \in \{1, 2, 4\}$. We use the monthly data in Welch and Goyal (2008), updated to 2019/12 and accessible on Amit Goyal's homepage. The data run from 1927/01 to 2019/12, with a total sample size of 1116. In terms of predictors, we use the ten economic variables in Welch and Goyal (2008), Rapach et al. (2009) and Liu and Kuo (2016) at time periods $t$, $(t-1)$ and $(t-2)$. These ten economic variables are the dividend price ratio, dividend yield, earnings price ratio, book-to-market ratio, net equity expansion, treasury bill, long-term return, default yield spread, default return spread and inflation; see Welch and Goyal (2008) for a detailed description of the data.

Since there are $30 = 10 \times 3$ variables in total, it is very time-consuming to use all potential models. Thus, we first sequence the variables and then use nested models as candidate models. Specifically, we first calculate the correlation coefficient between $EP_{t+h}$ and the ten economic variables at time period $t$ successively; then, we arrange these 10 variables in descending order of the absolute value of the correlation coefficient. Thus, we obtain a sequence of 30 nested models with the full sample, labeled from $\widetilde{m}_1$ to $\widetilde{m}_{30}$, based on predictors $\{x_{t,1}, x_{t-1,1}, x_{t-2,1}, \ldots, x_{t,10}, x_{t-1,10}, x_{t-2,10}\}$. Allowing for different window sizes, we add more candidate models $\widetilde{m}_{31}$–$\widetilde{m}_{120}$, among which models $\widetilde{m}_{31}$–$\widetilde{m}_{60}$ use the most recent 80% observations, $\widetilde{m}_{61}$–$\widetilde{m}_{90}$ use the most recent 60% and $\widetilde{m}_{91}$–$\widetilde{m}_{120}$ use the most recent 40%. We also use the model containing only the intercept term and using all the samples as a candidate model, denoted by $\widetilde{m}_0$. Apparently, the prediction under $\widetilde{m}_0$ is exactly HA. Let $\text{FVMA}_{31}$ and $\text{FVMA}_{121}$ represent our FVMA method with candidate models $\widetilde{m}_0$–$\widetilde{m}_{30}$ and $\widetilde{m}_0$–$\widetilde{m}_{120}$, respectively, where $121 = 1 + 30 \times 4$.

To evaluate the performance of all the methods, we use the observations from 1980/01 to 2019/12 as the test set. We compare all the methods via the squared forecast error, defined as

$$\text{SFE} = (EP_{t+h} - \widehat{EP}_{t+h})^2 - \widetilde{\sigma}^2_{t+h},$$

where $\widetilde{\sigma}^2_{t+h}$ is the estimator of the variance of $EP_{t+h}$. Here, $\widetilde{\sigma}^2_{t+h}$ is subtracted for the same reason as in (11). $\widehat{EP}_{t+h}$ is the $h$-step-ahead prediction of $EP_{t+h}$ and obtained using the data from periods 1 to $t$ as the training set. We use the in-sample mean-squared error of the $\text{FVMA}_{121}$ method as the estimator of the unknown variance, i.e., $\widetilde{\sigma}^2_{t+h}$. Table 3 lists the average and standard error of SFE for all methods.

### 7.2. Out-of-sample prediction results

To facilitate comparisons, we flag the best, second-best and third-best methods under each set of $h$ by the superscripts ①, ② and ③ in Table 3, respectively. Regardless of whether we consider one-step-ahead, two-step-ahead or four-step-ahead prediction, $\text{FVMA}_{121}$ outperforms the other approaches in terms of the average. Furthermore, our method shows the merit of robustness, since $\text{FVMA}_{121}$ and $\text{FVMA}_{31}$ show the best and second-best performance when $h = 1, 2$ and the best and third-best performance when $h = 4$. The HA and AdaBoost methods show obvious inferiority.

## 8. Discussions

As time series prediction follows temporal order and forward-validation preserves this essential characteristic, we propose a model averaging approach based on forward-validation. Moreover, our method allows the window size to vary in the candidate models, thereby simultaneously considering the uncertainty in the use of variables and sample size. When all the candidate models are misspecified, we prove that the prediction made by our method is asymptotically optimal in the sense that its corresponding prediction risk is asymptotically identical to the infimum of those from all the possible averaging predictions. If there exists one or more correctly specified candidate models, our method will automatically allocate all the weights to them.

In this paper, we treat the number of predictors in each candidate model as fixed. We believe our FVMA method is still applicable when this number increases with the sample sizes, but the establishment of the corresponding asymptotic properties needs further investigation. We set the parameter $\nu = T - 0.8T_{min}$ when applying the FVMA method. Although it does not affect the theoretical results of our method, it is still meaningful to explore how to make an optimal selection of $\nu$ in future research. In our empirical application, we sequence the variables for preparing the candidate models, which, in fact, is a model screening method. Claeskens et al. (2006) and Xie (2017) provided some alternative screening methods. By Theorem 2 of Zhang et al. (2016), the asymptotic optimality after model screening still holds under some regularity conditions. Last, it may be interesting to extend our FVMA method to predict quantiles, such as Lu and Su (2015) and Liao and Tsay (2020).

## Acknowledgments

We are very grateful to the editor, the associate editor, the two anonymous referees, and Changliang Zou for their very constructive comments and suggestions which substantially improved the original manuscript. All remaining errors are our own. Xinyu Zhang gratefully acknowledges research support from the National Natural Science Foundation of China (71925007, 72091212, 71988101 and 12288201), the CAS Project for Young Scientists in Basic Research (YSBR-008), and a joint grant from Academy for Multidisciplinary Studies, Capital Normal University, China.

## Appendix A. Proof of Theorem 1

Since $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{FV}(\mathbf{w})$ and $\sigma^2_{T+h}$ is a constant for $\mathbf{w}$, the FVMA model averaging weight

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \left\{\mathrm{FV}(\mathbf{w}) - \sigma^2_{T+h}\right\}.$$

To prove Theorem 1, by Lemma 1 in Zhang (2010) or Gao et al. (2019), it suffices to show that

$$\sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})} - 1 \right| = o(1) \tag{A.1}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{\mathrm{FV}(\mathbf{w}) - \sigma^2_{T+h} - R^*_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})} \right| = o_p(1). \tag{A.2}$$

First, we consider the verification of (A.1). Let $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}^{(1)\top}, \ldots, \widehat{\boldsymbol{\beta}}^{(M)\top})^\top$ and $\mathbf{B}^* = (\boldsymbol{\beta}^{*(1)\top}, \ldots, \boldsymbol{\beta}^{*(M)\top})^\top$. Then, we have

$$\frac{\partial \widehat{y}_{T+h}(\mathbf{w})}{\partial \widehat{\mathbf{B}}} = \sum_{j=1}^{M} w_j \frac{\partial \widehat{y}^{(j)}_{T+h}}{\partial \widehat{\mathbf{B}}} = \sum_{j=1}^{M} w_j \mathbf{I}_{(j)} (\mathbf{Y}^{(1)\top}_{T+h}, \ldots, \mathbf{Y}^{(M)\top}_{T+h})^\top \tag{A.3}$$

and

$$\frac{\partial \left\{y_{T+h} - \widehat{y}_{T+h}(\mathbf{w})\right\}^2}{\partial \widehat{\mathbf{B}}} = 2\left\{\sum_{i=1}^{M} w_i \widehat{y}^{(i)}_{T+h} - y_{T+h}\right\} \frac{\partial \widehat{y}_{T+h}(\mathbf{w})}{\partial \widehat{\mathbf{B}}}, \tag{A.4}$$

where $\mathbf{I}_{(j)}$ is a block diagonal matrix with the $j$th block being a $\dim(\boldsymbol{\beta}^{(j)})$-order identity matrix and the others being zeros. From (A.3), (A.4) and Assumptions 1–2, we can obtain that the components of the derivative $\partial \{\widehat{y}_{T+h}(\mathbf{w}) - y_{T+h}\}^2 / \partial \widehat{\mathbf{B}}|_{\widehat{\mathbf{B}}=\widetilde{\mathbf{B}}}$ are $O_p(1)$ uniformly for any $\widetilde{\mathbf{B}}$ between $\widehat{\mathbf{B}}$ and $\mathbf{B}^*$ and for any $\mathbf{w} \in \mathcal{H}$.

Let $L(\mathbf{w}) = \{y_{T+h} - \widehat{y}_{T+h}(\mathbf{w})\}^2 - \sigma^2_{T+h}$ and $L^*(\mathbf{w}) = \{y_{T+h} - y^*_{T+h}(\mathbf{w})\}^2 - \sigma^2_{T+h}$. From Assumption 3(i) and (iii), we obtain that

$$\begin{aligned}
&\sup_T \mathbb{E}\left\{\xi^{*-1}_{T+h} \max_{1\le i,j\le M} \left|\widehat{y}^{(i)}_{T+h} - y^{*(i)}_{T+h}\right| \left|\widehat{y}^{(j)}_{T+h} - y^{*(j)}_{T+h}\right|\right\}^{1+\delta_1/2} \\
\le &\sup_T \mathbb{E}\left\{\xi^{*-1}_{T+h} \max_{1\le i\le M} \left|\widehat{y}^{(i)}_{T+h} - y^{*(i)}_{T+h}\right|\right\}^{2+\delta_1} \left(\sup_T \xi^*_{T+h}\right)^{1+\delta_1/2} \\
< &\infty.
\end{aligned}$$

Thus, $\xi^{*-1}_{T+h} \max_{1\le i,j\le M} \left|\widehat{y}^{(i)}_{T+h} - y^{*(i)}_{T+h}\right| \left|\widehat{y}^{(j)}_{T+h} - y^{*(j)}_{T+h}\right|$ is uniformly integrable. We follow the symbol $e^{*(m)}_t = y_t - y^{*(m)}_t$ with $t \in \{\nu+1, \ldots, T+h\}$ and $m \in \{1, \ldots, M\}$ in Section 4. From Assumption 5(iii), we know that

$\xi_{T+h}^{*-1}\max_{1\le i,j\le M}\left|\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right|\left|e_{T+h}^{*(j)}\right|$ is uniformly integrable. Hence

$$\begin{aligned}
&\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|L(\mathbf{w})-L^*(\mathbf{w})\right| \\
=\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\{y_{T+h}-\widehat{y}_{T+h}(\mathbf{w})\}^2-\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right| \\
=\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\left\{\widehat{y}_{T+h}(\mathbf{w})-y_{T+h}^*(\mathbf{w})\right\}\left\{\widehat{y}_{T+h}(\mathbf{w})+y_{T+h}^*(\mathbf{w})-2y_{T+h}\right\}\right| \\
=\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\left\{\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right\}\left\{\widehat{y}_{T+h}^{(j)}+y_{T+h}^{*(j)}-2y_{T+h}\right\}\right| \\
\le\ &\xi_{T+h}^{*-1}\max_{1\le i\le M}\left|\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right|\max_{1\le j\le M}\left|\widehat{y}_{T+h}^{(j)}+y_{T+h}^{*(j)}-2y_{T+h}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left(\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\right) \\
=\ &\xi_{T+h}^{*-1}\max_{1\le i\le M}\left|\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right|\max_{1\le j\le M}\left|\widehat{y}_{T+h}^{(j)}-y_{T+h}^{*(j)}-2e_{T+h}^{*(j)}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left\{\left(\sum_{i=1}^{M}w_i\right)\left(\sum_{j=1}^{M}w_j\right)\right\} \\
\le\ &\xi_{T+h}^{*-1}\max_{1\le i,j\le M}\left|\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right|\left|\widehat{y}_{T+h}^{(j)}-y_{T+h}^{*(j)}\right| \\
&+2\xi_{T+h}^{*-1}\max_{1\le i,j\le M}\left|\widehat{y}_{T+h}^{(i)}-y_{T+h}^{*(i)}\right|\left|e_{T+h}^{*(j)}\right|,
\end{aligned}\tag{A.5}$$

implies that $\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}|L(\mathbf{w})-L^*(\mathbf{w})|$ is uniformly integrable. The last inequality in (A.5) can be satisfied because for any $\mathbf{w}\in\mathcal{H}$, we restrict the sum of all components of $\mathbf{w}$ to 1; see (7) for the definition of $\mathcal{H}$.

Based on Assumptions 1–2 and that the derivative $\partial\{y_{T+h}-\widehat{y}_{T+h}(\mathbf{w})\}^2/\partial\widehat{\mathbf{B}}|_{\widehat{\mathbf{B}}=\bar{\mathbf{B}}}=O_p(1)$ uniformly for any $\bar{\mathbf{B}}$ between $\widehat{\mathbf{B}}$ and $\mathbf{B}^*$ and for any $\mathbf{w}\in\mathcal{H}$, we can obtain that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}}\left|L(\mathbf{w})-L^*(\mathbf{w})\right| \\
=\ &\sup_{\mathbf{w}\in\mathcal{H}}\left|\{y_{T+h}-\widehat{y}_{T+h}(\mathbf{w})\}^2-\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right| \\
=\ &\sup_{\mathbf{w}\in\mathcal{H}}\left|\left(\widehat{\mathbf{B}}-\mathbf{B}^*\right)^\top\frac{\partial\{y_{T+h}-\widehat{y}_{T+h}(\mathbf{w})\}^2}{\partial\widehat{\mathbf{B}}}|_{\widehat{\mathbf{B}}=\bar{\mathbf{B}}}\right| \\
=\ &O_p\left((T-\nu)^{-1/2}\right).
\end{aligned}\tag{A.6}$$

As a result of (A.6) and the uniform integrability of $\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}|L(\mathbf{w})-L^*(\mathbf{w})|$, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R_{T+h}(\mathbf{w})}{R_{T+h}^*(\mathbf{w})}-1\right| \\
\le\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\mathbb{E}L(\mathbf{w})-\mathbb{E}L^*(\mathbf{w})\right| \\
\le\ &\mathbb{E}\left\{\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|L(\mathbf{w})-L^*(\mathbf{w})\right|\right\} \\
=\ &O\left(\xi_{T+h}^{*-1}(T-\nu)^{-1/2}\right).
\end{aligned}\tag{A.7}$$

Consequently, by (A.7) and Assumption 3(ii), (A.1) is obtained.

Next, for (A.2), we observe that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{FV}(\mathbf{w})-\sigma_{T+h}^2-R_{T+h}^*(\mathbf{w})}{R_{T+h}^*(\mathbf{w})}\right| \\
\le\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\mathrm{FV}(\mathbf{w})-\sigma_{T+h}^2-R_{T+h}^*(\mathbf{w})\right| \\
=\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\{y_t-\widetilde{y}_t(\mathbf{w})\}^2-\mathbb{E}\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right| \\
\le\ &\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\{y_t-\widetilde{y}_t(\mathbf{w})\}^2-\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|
\end{aligned}$$

$$+\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|$$
$$+\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right|. \tag{A.8}$$

Consequently, as long as the following three equations

$$\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\{y_t-\widetilde{y}_t(\mathbf{w})\}^2-\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|=o_p(1), \tag{A.9}$$

$$\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|=o_p(1) \tag{A.10}$$

and

$$\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right|=o(1) \tag{A.11}$$

can be verified, the proof of (A.2) is complete.

To prove (A.9), we denote the estimation of $\boldsymbol{\beta}^{(m)}$ built on the training set $\mathcal{T}_t^{(m)}$ as $\widetilde{\boldsymbol{\beta}}_t^{(m)}$; thus $\widetilde{y}_t^{(m)}=\mathbf{Y}_t^{(m)\top}\widetilde{\boldsymbol{\beta}}_t^{(m)}$, where $m\in\{1,\ldots,M\}$. Since the training set $\mathcal{T}_t^{(m)}$ excludes all of the samples observed later than time period $t$, $\mathcal{T}_t^{(m)}$ will not change when the full sample size $T$ increases, as well as $\widetilde{\boldsymbol{\beta}}_t^{(m)}$. Let $\widetilde{\mathbf{B}}_t=(\widetilde{\boldsymbol{\beta}}_t^{(1)\top},\ldots,\widetilde{\boldsymbol{\beta}}_t^{(M)\top})^\top$, where $t\in\{\nu+1,\ldots,T\}$. Note that $\widetilde{\mathbf{B}}_t=\widehat{\mathbf{B}}$ when the full sample size is $t$, hence, for any $m\in\{1,\ldots,M\}$, the sequence $\{\widetilde{\boldsymbol{\beta}}_t^{(m)}\}_{t=\nu+1}^{\infty}$ is a subsequence of $\{\widehat{\boldsymbol{\beta}}^{(m)}\}_T$, where $T$ goes to $\infty$. As $\widetilde{\mathbf{B}}_t=\widehat{\mathbf{B}}$ when the full sample size is $t$, from Assumption 1, we know that for any $\epsilon>0$, there exists a constant $C_\epsilon$ such that

$$\sup_T\Pr\left\{\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_T^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|\ge C_\epsilon(T-\nu)^{-1/2}\right\}<\epsilon. \tag{A.12}$$

We note that

$$\begin{aligned}&\sup_T\Pr\left\{\max_{t\in\{\nu+1,\ldots,T\}}\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_t^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|\ge C_\epsilon(T-\nu)^{-1/2}\right\}\\ &=\sup_T\Pr\left\{\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_{t_T}^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|\ge C_\epsilon(T-\nu)^{-1/2}\right\},\end{aligned} \tag{A.13}$$

where $\{t_T\}$ is a subsequence of $\{T\}$. Furthermore, based on the analysis following Assumption 1 and $(\nu+1)\le t_T\le T$, we have that $t_T=O(T)$, which along with the fact $\{t_T\}$ is a subsequence of $\{T\}$, implies that

$$\begin{aligned}&\sup_T\Pr\left\{\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_{t_T}^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|\ge C_\epsilon(T-\nu)^{-1/2}\right\}\\ &\le\sup_T\Pr\left\{\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_T^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|\ge C_\epsilon(T-\nu)^{-1/2}\right\}.\end{aligned} \tag{A.14}$$

Now, by (A.12)–(A.14), we obtain that

$$\max_{t\in\{\nu+1,\ldots,T\}}\max_{1\le m\le M}\left\|\widetilde{\boldsymbol{\beta}}_t^{(m)}-\boldsymbol{\beta}^{*(m)}\right\|=O_p((T-\nu)^{-1/2}). \tag{A.15}$$

Similar to (A.3)–(A.4), we can obtain that

$$\frac{\partial\widetilde{y}_t(\mathbf{w})}{\partial\widetilde{\mathbf{B}}_t}=\sum_{j=1}^{M}w_j\frac{\partial\widetilde{y}_t^{(j)}}{\partial\widetilde{\mathbf{B}}_t}=\sum_{j=1}^{M}w_j\mathbf{I}_{(j)}(\mathbf{Y}_t^{(1)\top},\ldots,\mathbf{Y}_t^{(M)\top})^\top \tag{A.16}$$

and

$$\frac{\partial\left\{y_t-\widetilde{y}_t(\mathbf{w})\right\}^2}{\partial\widetilde{\mathbf{B}}_t}=2\left\{\sum_{i=1}^{M}w_i\widetilde{y}_t^{(i)}-y_t\right\}\frac{\partial\widetilde{y}_t(\mathbf{w})}{\partial\widetilde{\mathbf{B}}_t}. \tag{A.17}$$

On the basis of (A.15)–(A.17) and Assumption 2, it can be guaranteed that the components of the derivative $\partial\left\{y_t-\widetilde{y}_t(\mathbf{w})\right\}^2/\partial\widetilde{\mathbf{B}}_t|_{\widetilde{\mathbf{B}}_t=\bar{\bar{\mathbf{B}}}_t}$ are $O_p(1)$ uniformly for any $\bar{\bar{\mathbf{B}}}_t$ between $\widetilde{\mathbf{B}}_t$ and $\mathbf{B}^*$ and for any $\mathbf{w}\in\mathcal{H}$ and $t\in\{\nu+1,\ldots,T\}$.

Therefore, similar to (A.6), from (A.15), we can obtain that

$$\begin{aligned}
&\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\{y_t-\widetilde{y}_t(\mathbf{w})\}^2-\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|\\
&=\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left(\widetilde{\mathbf{B}}_t-\mathbf{B}^*\right)^{\top}\frac{\partial\{y_t-\widetilde{y}_t(\mathbf{w})\}^2}{\partial\widetilde{\mathbf{B}}_t}|_{\widetilde{\mathbf{B}}_t=\bar{\widetilde{\mathbf{B}}}_t}\right|\\
&=O_p\left(\xi_{T+h}^{*-1}(T-\nu)^{-1/2}\right).
\end{aligned}\tag{A.18}$$

Thus, (A.9) follows from (A.18) and Assumption 3(ii).

To verify (A.10), we have that

$$\begin{aligned}
&\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2\right]\right|\\
&=\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\left\{e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_t^{*(i)}e_t^{*(j)}\right\}\right|\\
&\le\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left\{e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_t^{*(i)}e_t^{*(j)}\right\}\right|\\
&\le\xi_{T+h}^{*-1}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left\{e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_t^{*(i)}e_t^{*(j)}\right\}\right|\\
&=\frac{1}{\xi_{T+h}^{*}\sqrt{T-\nu}}\sum_{i=1}^{M}\sum_{j=1}^{M}\Psi_T(i,j),
\end{aligned}\tag{A.19}$$

where

$$\Psi_T(i,j)=\left|\frac{1}{\sqrt{T-\nu}}\sum_{t=\nu+1}^{T}\left\{e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_t^{*(i)}e_t^{*(j)}\right\}\right|.$$

By Theorem 3.49 and 5.20 in White (1984), Assumption 4 and Assumption 5 (i)–(ii), we obtain $\Psi_T(i,j)=O_p(1)$, for any $i,j\in\{1,\dots,M\}$. Therefore,

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\Psi_T(i,j)=O_p(1)\tag{A.20}$$

with $M$ is fixed. By (A.19), (A.20) and Assumption 3(ii), we obtain (A.10).

Last, we consider (A.11). We observe that

$$\begin{aligned}
&\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}\left\{y_t-y_t^*(\mathbf{w})\right\}^2-\mathbb{E}\left\{y_{T+h}-y_{T+h}^*(\mathbf{w})\right\}^2\right|\\
&\le\xi_{T+h}^{*-1}\sup_{\mathbf{w}\in\mathcal{H}}\sum_{i=1}^{M}\sum_{j=1}^{M}w_iw_j\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_{T+h}^{*(i)}e_{T+h}^{*(j)}\right|\\
&\le\xi_{T+h}^{*-1}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}e_t^{*(i)}e_t^{*(j)}-\mathbb{E}e_{T+h}^{*(i)}e_{T+h}^{*(j)}\right|\\
&=O\left(\xi_{T+h}^{*-1}(T-\nu)^{-1/2}\right),
\end{aligned}\tag{A.21}$$

where the last step is guaranteed by Assumption 5(iv). Thus, with (A.21) and Assumption 3(ii), (A.11) is verified. Hence, we have completed the proof of Theorem 1. ■

## Appendix B. Proof of Theorem 2

To prove Theorem 2, first we decompose FV(**w**) as follows:

$$\begin{aligned}\mathrm{FV}(\mathbf{w}) = R^*_{T+h}(\mathbf{w}) &+ \sigma^2_{T+h} + \frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\{y_t - \widetilde{y}_t(\mathbf{w})\}^2 - \left\{y_t - y^*_t(\mathbf{w})\right\}^2\right] \\ &+ \frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\left\{y_t - y^*_t(\mathbf{w})\right\}^2 - \mathbb{E}\left\{y_t - y^*_t(\mathbf{w})\right\}^2\right] \\ &+ \frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}\left\{y_t - y^*_t(\mathbf{w})\right\}^2 - \mathbb{E}\left\{y_{T+h} - y^*_{T+h}(\mathbf{w})\right\}^2. \end{aligned} \tag{B.1}$$

By (A.18)–(A.21), we can obtain

$$\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\{y_t - \widetilde{y}_t(\mathbf{w})\}^2 - \left\{y_t - y^*_t(\mathbf{w})\right\}^2\right] = O_p\left((T-\nu)^{-1/2}\right), \tag{B.2}$$

$$\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\left[\left\{y_t - y^*_t(\mathbf{w})\right\}^2 - \mathbb{E}\left\{y_t - y^*_t(\mathbf{w})\right\}^2\right] = O_p\left((T-\nu)^{-1/2}\right) \tag{B.3}$$

and

$$\frac{1}{T-\nu}\sum_{t=\nu+1}^{T}\mathbb{E}\left\{y_t - y^*_t(\mathbf{w})\right\}^2 - \mathbb{E}\left\{y_{T+h} - y^*_{T+h}(\mathbf{w})\right\}^2 = O\left((T-\nu)^{-1/2}\right). \tag{B.4}$$

Thus, from (B.1)–(B.4), the relationship between FV(**w**) and $R^*_{T+h}(\mathbf{w})$ can be written as

$$\mathrm{FV}(\mathbf{w}) = R^*_{T+h}(\mathbf{w}) + \sigma^2_{T+h} + O_p\left((T-\nu)^{-1/2}\right). \tag{B.5}$$

For any $j \in \mathcal{D}$, we have

$$\mathbb{E}\left\{y_{T+h} - y^{*(j)}_{T+h}\right\}^2 = \sigma^2_{T+h}. \tag{B.6}$$

Since there exists one or more correctly specified candidate models, we assume the $j_0$th is correct, so $j_0 \in \mathcal{D}$. Denote $\mathbf{w}_0^{(j_0)} \in \mathcal{H}$ as a vector with the $j_0$th element taking a value of 1 and all others taking a value of 0. Consequently, by (B.5) and (B.6), we obtain

$$\mathrm{FV}(\mathbf{w}_0^{(j_0)}) = \sigma^2_{T+h} + O_p\left((T-\nu)^{-1/2}\right). \tag{B.7}$$

As $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{FV}(\mathbf{w})$,

$$\mathrm{FV}(\widehat{\mathbf{w}}) \le \mathrm{FV}(\mathbf{w}_0^{(j_0)}) \tag{B.8}$$

is valid almost surely. By (B.5) and (B.7), we rewrite (B.8) as

$$R^*_{T+h}(\widehat{\mathbf{w}}) + O_p\left((T-\nu)^{-1/2}\right) \le O_p\left((T-\nu)^{-1/2}\right), \tag{B.9}$$

almost surely.

Now let $\widetilde{\mathbf{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_M)$ with $\widetilde{w}_i = 0$, $i \in \mathcal{D}$ and $\widetilde{w}_j = \dfrac{w_j}{1-\tau(\mathbf{w})}$, $j \notin \mathcal{D}$, so the relationship between $R^*_{T+h}(\mathbf{w})$ and $R^*_{T+h}(\widetilde{\mathbf{w}})$ can be written as

$$\begin{aligned} R^*_{T+h}(\mathbf{w}) &= \mathbb{E}\left\{y_{T+h} - y^*_{T+h}(\mathbf{w})\right\}^2 - \sigma^2_{T+h} \\ &= \mathbb{E}\left[\sum_{j\notin\mathcal{D}} w_j\left\{y_{T+h} - e_{T+h} - y^{*(j)}_{T+h}\right\}\right]^2 \\ &= \{1-\tau(\mathbf{w})\}^2\left(\mathbb{E}\left[\sum_{j\notin\mathcal{D}}\frac{w_j}{1-\tau(\mathbf{w})}\left\{y_{T+h} - e_{T+h} - y^{*(j)}_{T+h}\right\}\right]^2\right) \\ &= \{1-\tau(\mathbf{w})\}^2 R^*_{T+h}(\widetilde{\mathbf{w}}). \end{aligned} \tag{B.10}$$

As $\widetilde{\mathbf{w}}$ is defined corresponding to $\mathbf{w}$, we define $\widetilde{\widehat{\mathbf{w}}}$ corresponding to $\widehat{\mathbf{w}}$ in the same way. Thus, it is almost surely that

$$\begin{aligned}
&\{1-\tau(\widehat{\mathbf{w}})\}^2\,\widetilde{\xi}^*_{T+h}+O_p\left((T-\nu)^{-1/2}\right)\\
&\leq \{1-\tau(\widehat{\mathbf{w}})\}^2\,R^*_{T+h}(\widetilde{\widehat{\mathbf{w}}})+O_p\left((T-\nu)^{-1/2}\right)\\
&= R^*_{T+h}(\widehat{\mathbf{w}})+O_p\left((T-\nu)^{-1/2}\right)\\
&\leq O_p\left((T-\nu)^{-1/2}\right), \qquad \text{(B.11)}
\end{aligned}$$

where the equality results from (B.10) and the last inequality is guaranteed by (B.9). By (B.11) and Assumption 6, Theorem 2 is proved. ■

## Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.03.010.

## References


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

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

Cerqueira, V., Torgo, L., Smailović, J., Mozetič, I., 2017. A comparative study of performance estimation methods for time series forecasting. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 529–538.

Chen, J., Li, D., Linton, O., 2019. A new semiparametric estimation approach for large dynamic covariance matrices with multiple conditioning variables. J. Econometrics 212, 155–176.

Cheng, X., Hansen, B.E., 2015. Forecasting with factor–augmented regression: A frequentist model averaging approach. J. Econometrics 186, 280–293.

Cheng, T.-C.F., Ing, C.-K., Yu, S.-H., 2015. Toward optimal model averaging in regression models with time series errors. J. Econometrics 189, 321–334.

Cheng, X., Liao, Z., Shi, R., 2019. On uniform asymptotic risk of averaging GMM estimators. Quant. Econ. 10, 931–979.

Claeskens, G., Croux, C., Van Kerckhoven, J., 2006. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics 64, 972–979.

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

Falessi, D., Huang, J., Narayana, L., Thai, J.F., Turhan, B., 2020. On the need of preserving order of data when validating within-project defect classifiers. Empir. Softw. Eng. 25, 4805–4830.

Gao, Y., Zhang, X., Wang, S., Chong, T.T.-l., Zou, G., 2019. Frequentist model averaging for threshold models. Ann. Inst. Statist. Math. 71, 275–306.

Gao, Y., Zhang, X., Wang, S., Zou, G., 2016. Model averaging based on leave-subject-out cross-validation. J. Econometrics 192, 139–151.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

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

Hansen, B.E., Racine, J., 2012. Jackknife model averaging. J. Econometrics 167, 38–46.

Hao, H., Huang, B., Lee, T.-H., Unpublished results. Model averaging estimation of panel data models with many instruments and boosting.

He, B., Liu, Y., Wu, Y., Yin, G., Zhao, X., 2020. Functional martingale residual process for high-dimensional cox regression with model averaging. J. Mach. Learn. Res. 21, 1–37.

Heiler, P., Mareckova, J., 2021. Shrinkage for categorical regressors. J. Econometrics 223, 161–189.

Heyde, C.C., 1975. On the central limit theorem and iterated logarithm law for stationary process. Bull. Aust. Math. Soc. 12, 1–8.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Amer. Statist. Assoc. 98, 879–899.

Hjort, N.L., Claeskens, G., 2006. Focused information criteria and model averaging for the cox hazard regression model. J. Amer. Statist. Assoc. 101, 1449–1464.

Hjorth, U., 1982. Model selection and forward validation. Scand. J. Stat. 9, 95–105.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statist. Sci. 14, 382–417.

Isaksson, A., Shang, C., Sickles, R.C., 2021. Nonstructural analysis of productivity growth for the industrialized countries: A jackknife model averaging approach. Econometric Rev. 40, 321–358.

Jungmittag, A., 2016. Combination of forecasts across estimation windows: An application to air travel demand. J. Forecast. 35, 373–380.

Kaastra, I., Boyd, M., 1996. Designing a neural network for forecasting financial and economic time series. Neurocomputing 10, 215–236.

Leung, G., Barron, A.R., 2006. Information theory and mixing least–squares regressions. IEEE Trans. Inform. Theory 52, 3396–3410.

Liao, J.-C., Tsay, W.-J., 2020. Optimal multistep VAR forecast averaging. Econom. Theory 36, 1–28.

Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186, 142–159.

Liu, C.-A., Kuo, B.-S., 2016. Model averaging in predictive regressions. Econom. J. 19, 203–231.

Liu, Q., Okui, R., 2013. Heteroskedasticity-robust $C_p$ model averaging. Econom. J. 16, 463–472.

Liu, Q., Yao, Q., Zhao, G., 2020. Model averaging estimation for conditional volatility models with an application to stock market volatility forecast. J. Forecast. 39, 841–863.

Lohmeyer, J., Palm, F., Reuvers, H., Urbain, J.-P., 2019. Focused information criterion for locally misspecified vector autoregressive models. Econometric Rev. 38, 763–792.

Lu, X., Su, L., 2015. Jackknife model averaging for quantile regressions. J. Econometrics 188, 40–58.

Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. J. Econometrics 154, 139–153.

Magnus, J.R., Wan, A.T.K., Zhang, X., 2011. Weighted average least squares estimation with nonspherical disturbances and an application to the hong kong housing market. Comput. Stat. Data Anal. 55, 1331–1341.

McLeish, D.L., 1974. Dependent central limit theorems and invariance principles. Ann. Probab. 2, 620–628.

Ng, S., 2013. Chapter 14 – Variable selection in predictive regressions. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. In: Handbook of Economic Forecasting, vol. 2, Elsevier, pp. 752–789.

Pesaran, M.H., Pick, A., 2011. Forecast combination across estimation windows. J. Bus. Econom. Statist. 29, 307–318.

Pesaran, M.H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. J. Econometrics 137, 134–161.

Rapach, D.E., Strauss, J.K., Zhou, G., 2009. Out–of–sample equity premium prediction: Combination forecasts and links to the real economy. Rev. Financ. Stud. 23, 821–862.

Schnaubelt, M., 2019. A comparison of machine learning model validation schemes for non-stationary time series data. Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics, Nürnberg, FAU Discussion Papers in Economics, No. 11/2019.
Schönfeld, P., 1971. A useful central limit theorem for m-dependent variables. Metrika 17, 116–128.
Scott, D.J., 1973. Central limit theorems for martingales and for processes with stationary increments using a skorokhod representation approach. Adv. Appl. Probab. 5, 119–137.
Steel, M.F.J., 2020. Model averaging and its use in economics. J. Econ. Lit. 58, 644–719.
Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: An analysis and review. Int. J. Forecast. 16, 437–450.
Tu, Y., Wang, S., 2020. Jackknife model averaging for expectile regressions in increasing dimension. Econom. Lett. 197, 109607.
Ullah, A., Wang, H., 2013. Parametric and nonparametric frequentist model selection and model averaging. Econometrics 1, 157–179.
Wan, A.T.K., Zhang, X., Zou, G., 2010. Least squares model averaging by mallows criterion. J. Econometrics 156, 277–283.
Wang, Y., Hao, X., Wu, C., 2021. Forecasting stock returns: A time-dependent weighted least squares approach. J. Financial Mark. 53, 100568.
Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21, 1455–1508.
West, K.D., 2006. Chapter 3 – Forecast evaluation. In: Elliott, G., Granger, C., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Vol. 1. Elsevier, pp. 99–134.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.
White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, Orlando.
Wooldridge, J.M., White, H., 1988. Some invariance principles and central limit theorems for dependent heterogenrous processes. Econom. Therory 4, 210–230.
Xie, T., 2017. Heteroscedasticity-robust model screening: A useful toolkit for model averaging in big data analytics. Econom. Lett. 151, 119–122.
Xu, J., Yue, M., Zhang, W., 2020. A new multilevel modeling approach for clustered survival data. Econom. Theory 36, 707–750.
Yang, Y., 2001. Adaptive regression by mixing. J. Amer. Statist. Assoc. 96, 574–588.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: When and how? J. Amer. Statist. Assoc. 100, 1202–1214.
Zhang, X., 2010. Model Averaging and its Applications (Ph.D. Thesis). Academy of Mathematics and Systems Science, Chinese Academy of Sciences.
Zhang, X., 2021. A new study on asymptotic optimality of least squares model averaging. Econom. Theory 37, 388–407.
Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Ann. Statist. 39, 174–200.
Zhang, X., Lu, Z., Zou, G., 2013. Adaptively combined forecasting for discrete response time series. J. Econometrics 176, 80–91.
Zhang, X., Yu, D., Zou, G., Liang, H., 2016. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. J. Amer. Statist. Assoc. 111, 1775–1790.