State-Varying Model Averaging Prediction*

Yuying Sun
State Key Laboratory of Mathematical Sciences
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
Shaoxin Hong†
Center for Economic Research, Shandong University, Jinan 250100, Shandong, China
Zongwu Cai
Department of Economics, University of Kansas, Lawrence, KS 66045, USA.

March 25, 2025

Abstract

This paper proposes a novel state-varying model averaging prediction for varying-coefficient models that accounts for parameter uncertainty and model misspecification. We develop a leave-h-out state-dependent forward-validation criterion to select state-varying combination weights. It is shown that the proposed averaging prediction is asymptotically optimal in the sense of achieving the lowest possible out-of-sample prediction risk in a class of model averaging estimators. This complements existing model averaging methods that primarily focus on minimizing the in-sample squared error loss. Besides, when the set of candidate models includes correctly specified models, the proposed approach asymptotically assigns full weight to these models. Furthermore, the proposed approach is flexible and encompasses special cases including ultra-high dimensional models as well as state-varying factor-augmented regression models. Simulation studies and empirical applications highlight the merits of the proposed averaging prediction relative to other existing model averaging and model selection methods.

Keywords: Asymptotic optimality; Varying-coefficient models; Forward-validation; Model averaging; Weight convergence

*Sun was supported by the National Natural Science Foundation of China (72322016, 72073126, 72091212, 71988101), and Hong was supported by the National Natural Science Foundation of China (72033008, 72133002, and 72303131).

†Corresponding author.

1 Introduction

Extensive empirical evidence suggests that the structure of economic relationships evolves over state variables, driven by factors such as policy reforms, business cycle fluctuations, technological advancements, and shifts in agents' preferences. For instance, in economics, Jansen et al. (2008) found that the impact of monetary policy on stock markets fluctuates based on whether the fiscal stance is expansionary or contractionary, which is further explored by Tu and Wang (2020). The business cycle serves as another well-known illustration, highlighting how the dynamic properties of economic variables can differ between expansionary and recessionary phases (Hamilton, 1989). In financial markets, stock return volatility exhibits a tendency to be higher during recessionary periods compared to expansionary phases (Paye, 2012). To capture these state-dependent features, varying (functional)-coefficient models, which allow coefficients to vary as functions of state variables, have gained popularity in statistics and economics as well as finance over the last three decades; see, for example, the survey paper by Cai (2010) and the references therein.

Recently, in a data-rich environment, a substantial number of varying-coefficient candidate models arise, so that model selection becomes a popular approach to identify the optimal model for prediction. Common model selection criteria include the nonparametric version of the bias-corrected Akaike information criterion (AICc, Cai and Tiwari (2000)) and the Bayesian information criterion (BIC, Lian (2012)). However, the selected model may omit valuable information from alternative models and ignore the uncertainty across different candidate models (Zhang and Zhang, 2023). Unlike model selection, model averaging incorporates all available information and constructs a weighted average of all potential candidate models, which serves as a form of insurance against selecting a poor candidate model

(Hansen, 2014; Hsiao and Wan, 2014). Existing frequentist averaging methods have received much attention in statistics and econometrics (Hansen, 2007; Li et al., 2022; Jiang et al., 2024), but typically consider unbiased loss risk estimation over the entire sample, yielding constant weights. However, the optimal predictive model may vary across different economic states, as the best forecast during an expansion period may lose its predictive ability during a recession. Thus, allowing data-driven weights to adapt over state variables becomes highly desirable. Intuitively, well-performing candidate models should receive larger weights, while underperforming ones should receive smaller or zero weights at different state stages.

To the best of our knowledge, only three papers by Sun et al. (2021, 2023) and Chen et al. (2024) considered to select optimal non-constant (time-varying) weights in the frequentist model averaging literature. Specifically, Sun et al. (2021) proposed time-varying model averaging estimators based on a local jackknife criterion, which was further studied by Sun et al. (2023) in a penalized setting and by Chen et al. (2024) in time-varying factor models. However, the aforementioned studies allow weights to vary over time, not state variables, and establish asymptotic properties based on an in-sample squared error loss rather than an out-of-sample prediction risk, thereby limiting their applicability for out-of-sample forecasting. Furthermore, while there exist compelling theoretical motivations for permitting dynamic combination weights, empirical evidence frequently indicates that a simple equal-weighted average of candidate forecasts achieves superior out-of-sample predictive performance relative to more intricate adaptive weighting schemes. This phenomenon is termed as the "forecast combination puzzle" in the literature (Stock and Watson, 2004; Claeskens et al., 2016).

Motivated by re-visiting this puzzle, we develop a novel model averaging prediction with state-varying weights in varying-coefficient models under the large dimensional framework in the sense that both the dimension of covariates and the number of candidate models are

allowed to be divergent. In such a way, we propose a state-varying model averaging (SVMA) prediction that selects state-dependent weights for varying-coefficient candidate models by minimizing a local forward-validation weight choice criterion. The proposed weight choice criterion is designed for selecting optimal weights in out-of-sample forecasts and is suitable for highly persistent time-series data. We establish the asymptotic optimality under the large dimensional framework. Importantly, these asymptotic properties of the out-of-sample prediction risk complement existing model averaging methods that primarily focus on the in-sample squared error loss function. Besides, when the correctly specified models are included among the candidate models, our proposed method assigns all weights to these correctly specified models. This novel result of asymptotically selecting the correctly specified models corresponds to the consistency property in model selection. Also, we derive asymptotic results showing that our SVMA procedure dominates the simple averaging (SA) approach in out-of-sample forecasting performance, thereby providing a novel perspective to resolve the forecast combination puzzle. We further extend our work to the ultra-high dimensional models as well as state-varying factor-augmented regression models, and the asymptotic optimality is investigated accordingly. Simulation studies and empirical applications highlight the merits of the proposed SVMA, relative to various popular estimators with constant model averaging weights and model selection.

The main contributions of this paper is summarized as follows. First, the proposed state-dependent forward-validation weight choice criterion in this paper varies over state variables. Consequently, the data-driven weights are conditioned on the state variable's outcome, providing insights for building economic models and improving forecast robustness. Unlike the existing frequentist strategies, such as Mallows-type criteria (Hansen, 2007; Zhu et al., 2019), cross-validation (Hansen and Racine, 2012), forward-validation (Zhang and Zhang,

2023), and semiparametric model averaging prediction (Li et al., 2018), which are designed to select constant weights, our approach allows weights to change dynamically over state variables. With the proposed weight choice, time-variation is captured in the state-varying weight structure as the state process evolves dynamically, including smooth and discontinuous processes. Thus, our approach covers cases of some large shifts or numerous gradual adjustments. Weights can fluctuate for particular state realizations or remain constant for other state process outcomes. As a result, we incorporate time-invariant model averaging as a special case of our SVMA estimator and extend it to time-variant model averaging with minor modifications.

Second, we establish the asymptotic optimality and consistency of the proposed SVMA based on the out-of-sample prediction risk, rather than the in-sample squared error loss commonly used in the existing time-varying model averaging studies (Sun et al., 2021, 2023; Chen et al., 2024). The asymptotic optimality from the prediction aspect is not a trivial extension of already existing results, because the frameworks developed in Li (1987), Hansen and Racine (2012), and Chen et al. (2024) cannot be directly applied. Rather than invoking Whittle's inequality, we provide a new strategy to bound the divergence between the local forward-validation and out-of-sample prediction risk function. Additionally, we provide a theoretical justification for the superior performance of our SVMA approach over SA, an aspect that has been unexplored in the existing literature on time-varying model averaging. Furthermore, the proposed approach encompasses special cases including ultra-high dimensional models as well as state-varying factor-augmented regression models. To the best of our knowledge, these findings on asymptotic optimality and consistency based on prediction risk are new to the model averaging literature.

Finally, from theoretical perspective, we overcome some difficulties in deriving the

asymptotic properties for state-varying weights and varying-coefficients in candidate models. For example, unlike constant model averaging (Hansen and Racine, 2012; Fang et al., 2019, 2023), projection matrices in each candidate model lack symmetry and idempotent. In contrast to time-varying model averaging (Chen et al., 2024), the data-driven combination weights in SVMA are deterministic functions of the state process, which is a stochastic process. Consequently, they are dynamic and conditional on the realized outcome of the state process. Also, allowing both covariate dimensionality and the number of candidate models to diverge complicates the proofs, such as the parameter consistency for misspecified models in Lemma 3 in Appendix A.

The rest of this paper is organized as follows. Section 2 proposes the state-dependent weight choice criterion, and derives the asymptotic properties. Section 3 discusses special cases, including the ultra-high dimensional models and the state-varying factor-augmented regression models. Sections 4 and 5 present simulation studies and two empirical examples, respectively. Finally, Section 6 concludes. All mathematical proofs are relegated to Appendices A and B.

2 Modeling Procedures

2.1 Model Setup

Let $\{\mathbf{U}_t, \mathbf{X}_t, Y_{t+h}\}_{t=1}^{\infty}$ be a jointly strictly stationary process with $\mathbf{U}_t$ taking values in $\mathbb{R}^p$ and $\mathbf{X}_t$ taking values in $\mathbb{R}^q$. The data generating process (DGP) is considered as follows:

$$Y_{t+h} = m(\mathbf{U}_t, \mathbf{X}_t) + \epsilon_{t+h} \equiv \mu_t + \epsilon_{t+h}, \quad t = 1, \ldots, T,$$

where Y_{t+h} is a dependent variable, $\mathbf{X}_t = (X_{t1}, X_{t2}, \ldots, X_{tq})'$ is a vector of covariate, $\mu_t = m(\mathbf{u}, \mathbf{x}) = \mathbb{E}(Y_{t+h}|\mathbf{U}_t = \mathbf{u}, \mathbf{X}_t = \mathbf{x})$ is the multivariate regression function, ϵ_{t+h} is the unobservable disturbance, and h is the forecast horizon. Here, both $\mathbf{X}_t$ and $\mathbf{U}_t$ might contain some lagged values of Y_{t+h}. For notational simplicity, let $\mathbf{Y} = (Y_{1+h}, \ldots, Y_T)'$ be a $(T-h) \times 1$ vector of the observed values of the dependent variable, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{T-h})'$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{T-h})'$ be a $(T-h) \times q$ covariate matrix, and $\boldsymbol{\epsilon} = (\epsilon_{1+h}, \ldots, \epsilon_T)'$. We follow Zhu et al. (2019) and Zhang and Zhang (2023) to set the innovation terms satisfying $\mathbb{E}(\epsilon_{t+h}|\mathbf{X}_t, \mathbf{U}_t) = 0$ and $\mathrm{Var}(\epsilon_{t+h}) = \sigma^2_{t+h}$.

Clearly, the nonparametric estimation of $m(\mathbf{u}, \mathbf{x})$ in $\mathbb{R}^{p+q}$ might suffer from the so-called *curse of dimensionality*. To overcome this difficulty, as argued in Cai et al. (2000), the varying-coefficient regression model has the particular form as

$$m(\mathbf{U}_t, \mathbf{X}_t) = \sum_{j=1}^{q} \alpha_j(\mathbf{U}_t) X_{tj} = \mathbf{X}_t' \boldsymbol{\alpha}(\mathbf{U}_t) = \mu_t, \tag{1}$$

where $\{\alpha_j(\cdot)\}_{j=1}^q$ are measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$, which is flexible enough to cover many applications. For this regard, the reader is referred to the survey paper by Cai (2010) for details on some examples particularly appealing in modeling economic and financial data. Here, to ease notation, it is assumed that $\mathbf{U}_t$ is an observable scalar smoothing variable. Therefore, in what it follows, it is assumed that $p = 1$ and $\mathbf{U}_t$ is changed to U_t. Furthermore, the varying-coefficient form in (1) can be regarded as an approximation of $m(U_t, \mathbf{X}_t)$. For convenience, it is assumed that $\mathbf{X}_t$ is a scalar. Then, by Taylor expansion and assuming that $m(u, x)$ is differentiable with respect to x in infinite order, it is easy to obtain that $m(u, x) = \sum_{j=0}^{\infty} \frac{1}{j!} \frac{\partial^j m(u,x)}{\partial x^j}|_{x=0} x^j = \sum_{j=0}^{\infty} a_j(u)\, z_j \approx \sum_{j=0}^{q} a_j(u)\, z_j$ for some q, where $z_j = x^j$ for all j. Therefore, Y_{t+h} can be approximated by $\mathbf{X}_t' \boldsymbol{\alpha}(U_t) + \epsilon_{t+h}$, which is in the form of model (1).

Now, we use M candidate models to approximate the true regression function $m(\cdot,\cdot)$, where M is allowed to depend on the sample size T. The m-th candidate model is given by

$$Y_{t+h} = \sum_{j=1}^{q_m} \alpha_j^{(m)}(U_t) X_{tj} + \epsilon_{t+h}^{(m)} \equiv \mathbf{X}_t^{(m)\prime} \boldsymbol{\alpha}^{(m)}(U_t) + \epsilon_{t+h}^{(m)} = \mu_t^{(m)} + \epsilon_{t+h}^{(m)},$$

where the functions $\{\alpha_j^{(m)}(\cdot)\}_{j=1}^{q_m}$ are measurable functions from $\mathbb{R}$ to $\mathbb{R}$ and $\mathbf{X}_t^{(m)}$ is a $q_m \times 1$ vector of regressors and $\boldsymbol{\alpha}^{(m)}(U_t) = (\alpha_1^{(m)}(U_t), \ldots, \alpha_{q_m}^{(m)}(U_t))'$. Note that each candidate model is allowed to have a divergent dimension of regressors as the sample size T increases; that is, q_m grows to infinity at some slower rate than the sample size, and all candidate models might be nested and/or non-nested, i.e., $\mathbf{X}_t^{(m)} = \mathbf{\Pi}^{(m)} \mathbf{X}_t$, where $\mathbf{\Pi}^{(m)}$ is an appropriate matrix of size $q_m \times q$ mapping $\mathbf{X}_t$ to $\mathbf{X}_t^{(m)}$. For example, if all candidate models are nested, then, $\mathbf{\Pi}^{(m)} = (\mathbf{I}_{q_m}, \mathbf{0}_{q_m \times (q-q_m)})$ and $\mathbf{I}_{q_m}$ is the $q_m \times q_m$ identity matrix.

2.2 Estimation Procedure

For ease of notation, the unknown coefficient parameters can be estimated by using a local constant estimation technique[1]. For any given u_0 and U_t in a neighborhood of u_0, it follows from the Taylor expansion that $\alpha_j^{(m)}(U_t) = \alpha_j^{(m)}(u_0) + O_p(U_t - u_0)$, where $\alpha_j^{(m)}(u_0)$ is the local intercept corresponding to $\alpha_j^{(m)}(U_t)$. Using the data with U_t around u_0, we run the following local constant regression. Minimizing with respect to $\{\alpha_j^{(m)}(u_0)\}$, we have the locally weighted sum squared errors

$$\sum_{t=1}^{T-h} \left[Y_{t+h} - \sum_{j=1}^{q_m} \alpha_j^{(m)}(u_0) X_{tj} \right]^2 k_t, \tag{2}$$

where $k_t = k((U_t - u_0)/l)$, $k(\cdot)$ is a kernel function on $\mathbb{R}$, and $l > 0$ is a bandwidth which satisfies $l \to 0$ and $Tl \to \infty$. Let $\mathbf{X}^{(m)}$ denote a $(T-h) \times q_m$ matrix with $\mathbf{X}_t^{(m)\prime}$ as

[1] Of course, one can use a local polynomial estimation tool as in Cai et al. (2000).

its t-th row and $\mathbf{K}(u_0) = \text{diag}\{k_1, \ldots, k_{T-h}\}$. Then, the locally weighted least squared errors in (2) can be rewritten as $(\mathbf{Y} - \mathbf{X}^{(m)}\boldsymbol{\alpha}^{(m)}(u_0))'\mathbf{K}(u_0)(\mathbf{Y} - \mathbf{X}^{(m)}\boldsymbol{\alpha}^{(m)}(u_0))$, where $\boldsymbol{\alpha}^{(m)}(u_0) \equiv (\alpha_1^{(m)}(u_0), \ldots, \alpha_{q_m}^{(m)}(u_0))'$. Thus, the local constant estimator of $\boldsymbol{\alpha}^{(m)}(u_0)$ is given by $\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) = \left[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}\right]^{-1}\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{Y}$ and $\widehat{\alpha}_j^{(m)}(u_0) = \mathbf{e}'_{j,q_m}\widehat{\boldsymbol{\alpha}}^{(m)}(u_0)$ with $\mathbf{e}_{j,q_m}$ the $q_m \times 1$ unit vector with 1 at the j-th position. Define $\mathbf{P}_t^{(m)} \equiv \mathbf{P}_t^{(m)}(U_t) = \left[\mathbf{X}^{(m)'}\mathbf{K}(U_t)\mathbf{X}^{(m)}\right]^{-1}\mathbf{X}^{(m)'}\mathbf{K}(U_t)$, a $q_m \times (T-h)$ matrix. Then, $\widehat{\boldsymbol{\alpha}}^{(m)}(U_t) = \mathbf{P}_t^{(m)}\mathbf{Y}$ and the least square estimation $\widehat{\boldsymbol{\mu}}^{(m)}$ for the conditional mean in the m-th candidate model is $\widehat{\boldsymbol{\mu}}^{(m)} = (\widehat{\mu}^{(m)}(U_1), \cdots, \widehat{\mu}^{(m)}(U_{T-h}))' = (\mathbf{P}_1^{(m)'}\mathbf{X}_1^{(m)}, \cdots, \mathbf{P}_{T-h}^{(m)'}\mathbf{X}_{T-h}^{(m)})'\mathbf{Y} \equiv \mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y}$. The h-step-ahead prediction of Y_{T+h} in the m-th candidate model is $\widehat{Y}_{T+h}^{(m)} = \mathbf{X}_T^{(m)'}\widehat{\boldsymbol{\alpha}}^{(m)}(U_T)$.

2.3 Selection of State-varying Weight

To simplify notation, we write $w^m(u_0)$ as w^m for a given u_0. Then, the weight vector $\mathbf{w} = (w^1, \ldots, w^M)' = \mathbf{w}(u_0)$ is a vector of weights in the unit simplex of $\mathbb{R}^M$, i.e., $\mathcal{H} = \left\{\mathbf{w} \in [0,1]^M : \sum_{m=1}^M w^m = 1\right\}$, depending on the grid point u_0. Actually, these weights can be allowed to be dynamic such that they can be state-varying weights of some information. Define $\mathbf{P}(\mathbf{w}, \mathbf{X}) = \sum_{m=1}^M w^m \mathbf{P}^{(m)}(\mathbf{X})$. For given $\mathbf{w}$, an averaging estimator for the conditional mean is given by

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{m=1}^M w^m \widehat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^M w^m \mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y} = \mathbf{P}(\mathbf{w}, \mathbf{X})\mathbf{Y},$$

and the model averaging estimator for $\boldsymbol{\alpha}(u_0)$ is given by $\widehat{\boldsymbol{\alpha}}(u_0, \mathbf{w}) = \sum_{m=1}^M w^m \boldsymbol{\Pi}^{(m)'}\widehat{\boldsymbol{\alpha}}^{(m)}(u_0)$, and the averaging prediction for Y_{T+h} is $\widehat{Y}_{T+h}(\mathbf{w}) = \sum_{m=1}^M w^m \widehat{Y}_{T+h}^{(m)}$.

Within heteroskedastic or autocorrelated errors, we propose the leave-h-out forward-validation estimator in the varying-coefficient linear regression model. Denote the following matrices $\boldsymbol{\phi}_t = (\mathbf{I}_t, \mathbf{0}_{t\times(T-h-t)})$ for $1 \le t \le h$, $\boldsymbol{\phi}_t = (\mathbf{0}_{h\times(t-h)}, \mathbf{I}_h, \mathbf{0}_{h\times(T-h-t)})$ for $h+1 \le t \le$

$T-h$, $\boldsymbol{\pi}_t = (\mathbf{0}_{1\times(t-1)}, 1)$ for $1 \le t \le h$, and $\boldsymbol{\pi}_t = (\mathbf{0}_{1\times(h-1)}, 1)$ for $h+1 \le t \le T-h$. Then, we obtain $\mathbf{Y}_{[t+h]} = \boldsymbol{\phi}_t \mathbf{Y}$ and $\mathbf{X}^{(m)}_{[t]} = \boldsymbol{\phi}_t \mathbf{X}^{(m)}$, which are the sets to be removed. Denote $\mathbf{Y}_{[-(t+h)]}$ and $\mathbf{X}^{(m)}_{[-t]}$ as the remaining sets of $\mathbf{Y}$ and $\mathbf{X}^{(m)}$ after removing $\mathbf{Y}_{[t+h]}$ and $\mathbf{X}^{(m)}_{[t]}$, respectively. For any given u_0, the local constant estimator of $\boldsymbol{\alpha}^{(m)}(u_0)$ is obtained as $\widehat{\boldsymbol{\alpha}}^{(m)}_{[-t]}(u_0) = \left(\mathbf{X}^{(m)\prime}_{[-t]}\mathbf{K}_{[-t]}(u_0)\mathbf{X}^{(m)}_{[-t]}\right)^{-1}\mathbf{X}^{(m)\prime}_{[-t]}\mathbf{K}_{[-t]}(u_0)\mathbf{Y}_{[-(t+h)]}$, and the leave-$h$-out forward-validation estimator $\widetilde{\mu}^{(m)}_t(u_0)$ of $\mu^{(m)}_t$ is

$$\widetilde{\mu}^{(m)}_t(u_0) = \boldsymbol{\pi}_t \mathbf{X}^{(m)}_{[t]} \widehat{\boldsymbol{\alpha}}^{(m)}_{[-t]}(u_0), \tag{3}$$

where $\mathbf{K}_{[-t]}(u_0) = \mathrm{diag}\{k_1, \ldots, k_{t-h}, k_{t+1}, \ldots, k_{T-h}\}$. Thus, the forward-validation averaging estimator of μ_t is $\widetilde{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M} w^m \widetilde{\mu}^{(m)}_t(u_0)$ and then $\widetilde{\boldsymbol{\mu}}(\mathbf{w}) = (\widetilde{\mu}_1(\mathbf{w}), \ldots, \widetilde{\mu}_{T-h}(\mathbf{w}))'$.

We propose the state-dependent forward-validation criterion to select the state-varying combination weights for the averaging prediction as follows:

$$\mathrm{SFV}(u_0, \mathbf{w}) = \frac{1}{(T-h)l}(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(u_0)(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})).$$

For any given u_0, minimizing $\mathrm{SFV}(u_0, \mathbf{w})$ with respect to $\mathbf{w}$, we have

$$\widehat{\mathbf{w}}(u_0) = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{SFV}(u_0, \mathbf{w}).$$

Then, the SVMA prediction for Y_{T+h} is $\widehat{Y}_{T+h}(\widehat{\mathbf{w}}(U_T)) = \sum_{m=1}^{M} \widehat{w}^m(U_T)\widehat{Y}^{(m)}_{T+h}$, where $\widehat{\mathbf{w}}(U_T) = (\widehat{w}^1(U_T), \ldots, \widehat{w}^M(U_T))'$. The detailed implementation for the proposed model averaging procedure can be formulated as a quadratic programming subject to linear constraints; see algorithm 1 in Appendix A.

2.4 Asymptotic Properties

In this section, we present the asymptotic properties of the proposed SVMA. To this end, define the risk function as $R_{T+h}(\mathbf{w}) = \mathbb{E}[Y_{T+h} - \widehat{Y}_{T+h}(\mathbf{w})]^2 - \sigma^2_{T+h}$. Note that unlike the

existing time-varying model averaging literature (Sun et al., 2021, 2023; Chen et al., 2024), the error variance σ^2_{T+h} is substracted because it is often the leading term in the mean-squared forecast error, which is consistent with the idea in Hansen (2008)[2]. Ideally, one would seek to select the combination weight $\mathbf{w}$ so as to minimize the out-of-sample prediction risk function $R_{T+h}(\mathbf{w})$ subject to the constraint set $\mathcal{H}$. However, this is infeasible due to the expectation depending on the unknown conditional probability density function. Rather than directly minimizing the infeasible risk $R_{T+h}(\mathbf{w})$, we select the data-driven weights by minimizing a state-dependent forward-validation criterion and demonstrate the selected state-varying weights asymptotically minimize the prediction risk function.

Let $\boldsymbol{\alpha}^{(m)*}(u_0)$ be parameter vector which is essentially derived from minimizing the MSE between Y_{t+h} and the m-th candidate model at the point u_0. From Lemma 3 in Appendix A, for any given u_0 and any $m \in \{1, \cdots, M\}$, we have

$$||\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0)|| = O_p(q^{1/2}(T-h)^{-1/2}l^{-1/2}).$$

Remark 1. *This is similar to that of parametric estimation in the existing literature. For example, based on Assumptions A1-A3(a) and A4-A6(a) of Theorem 3.2 in White (1982), the consistency of parameter estimator in misspecified models can be derived under the maximum likelihood framework. Besides, with the under-smoothing bandwidth, the squared bias term* $O_p(ql^4)$ *of* $\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0)$ *could be dominated by the variance term* $O_p(q(T-h)^{-1}l^{-1})$. *Thus, the bias term can be ignored.*

Define $R^*_{T+h}(\mathbf{w}) \equiv \mathbb{E}\{Y_{T+h} - Y^*_{T+h}(\mathbf{w})\}^2 - \sigma^2_{T+h}$, $Y^*_{T+h}(\mathbf{w}) \equiv \sum_{m=1}^M w^m Y^{(m)*}_{T+h}$, where $Y^{(m)*}_{T+h} \equiv \mathbf{X}_T^{(m)'}\boldsymbol{\alpha}^{(m)*}(U_T)$, and $\xi^*_{T+h}(U_T) = \inf_{\mathbf{w}\in\mathcal{H}} R^*_{T+h}(\mathbf{w})$. Let $\epsilon^{(m)*}_{t+h} = Y_{t+h} - Y^{(m)*}_{t+h}$, which

[2]To illustrate this, under some regularity conditions we decompose the mean squared forecast error of prediction as follows: $\mathbb{E}[Y_{T+h} - \widehat{Y}_{T+h}]^2 = \mathbb{E}[\epsilon_{T+h} + \mathbf{X}'_T(\boldsymbol{\alpha}(U_T) - \widehat{\boldsymbol{\alpha}}(U_T))]^2 = \sigma^2_{T+h} + \frac{1}{Tl}\mathbb{E}\left\{\mathbf{X}'_T\left(\frac{1}{Tl}\mathbf{X}'\mathbf{K}(U_T)\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{Tl}}\mathbf{X}'\mathbf{K}(U_T)\boldsymbol{\epsilon}\right)\right\}^2 = \sigma^2_{T+h} + o(1)$.

is interpreted as the bias of the m-th model in forecasting Y_{t+h} when $(T-h)l$ diverges. Again, denote the joint density of $(U, \mathbf{X})$ by $f(u, \mathbf{x})$ and the marginal density of U by $f_U(u)$. Finally, let $\zeta_{\max}(\mathbf{A})$ and $\zeta_{\min}(\mathbf{A})$ be the maximum and minimum singular values of a matrix $\mathbf{A}$, respectively. Unless stated otherwise, all limiting processes refer to $(T-h)l \to \infty$, where the forecast horizion h is fixed.

Condition (C.1). *For all $s \geq 1$ and some positive constant C, $|f(u, v|\mathbf{x}_0, \mathbf{x}_1; s)| \leq C < \infty$, and $f(u|\mathbf{x}) \leq C < \infty$, where $f(u, v|\mathbf{x}_0, \mathbf{x}_1; s)$ is the conditional density of (U_0, U_s) given $\mathbf{X}_0 = \mathbf{x}_0$ and $\mathbf{X}_s = \mathbf{x}_1$ and $f(u|\mathbf{x})$ is the conditional density of U given $\mathbf{X} = x$.*

Condition (C.2). *$\{U_t, \mathbf{X}_t, Y_{t+h}\}$ is α-mixing process with the mixing coefficient $\{\alpha(j)\}$ satisfying that $\sum j^c \alpha(j)^{1-2/\iota} < \infty$ for some $\iota > 2$ and $c > 1 - 2/\iota$, $\sup_{1 \leq t \leq T-h}(T-h)^{-1}l^{-1}||\sqrt{\mathbf{K}(u_0)}\boldsymbol{\mu}||^2 = O_p(1)$ for any given u_0, $\sup_{1 \leq t \leq T-h} ||\mathbf{X}_t^{(m)}||/\sqrt{q} = O_p(1)$ uniformly for all m, and $\zeta_{\min}((T-h)^{-1}l^{-1}\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}) \geq C_0$ for some positive constant C_0.*

Condition (C.3). *There exist positive constants c_1 and c_2 such that for any $i, j \in \{1, \ldots, M\}$ and given u_0 (i) $\mathrm{Var}\left((T-h)^{-1/2}l^{-1/2}\sum_{t=1}^{T-h} k_t \epsilon_{t+h}^{(i)*}\epsilon_{t+h}^{(j)*}\right) \geq c_1 > 0$ and (ii) $\mathrm{Var}\left\{\epsilon_{t+h}^{(i)*}\epsilon_{t+h}^{(j)*}\right\} \leq c_2 < \infty$, $1 \leq t \leq T$.*

Condition (C.4). *The kernel function $k : [-1, 1] \to \mathbb{R}^+$ is a bounded symmetric probability density function, satisfying that $\int_{-1}^{1} k(u)du = 1$, $\int_{-1}^{1} k^2(u)du < \infty$, and $\int_{-1}^{1} k(u)u^2 du < \infty$.*

Condition (C.5). *The bandwidth $l = cT^{-\nu}$ for some positive ν and $0 < c < \infty$.*

Condition (C.6). *$\xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w} \in \mathcal{H}}\{[Y_{T+h} - \widehat{Y}_{T+h}(\mathbf{w})]^2 - [Y_{T+h} - Y_{T+h}^*(\mathbf{w})]^2\}$ is uniformly integrable.*

Condition (C.7). *$qMT^{-1}l^{-1} = o(1)$, $qMT^{-1/2}l^{-1/2}\xi_{T+h}^{*-1}(U_T) = o(1)$, $M^2T^{-1/2}l^{-1/2}\xi_{T+h}^{*-1}(U_T) = o(1)$ and $q^2MT^{-1}l^{-1}\xi_{T+h}^{*-1}(U_T) = o(1)$.*

Remark 2. *Condition (C.1) is the same as Condition 1(ii) in Cai et al. (2000), and Condition (C.2) imposes a standard requirement on the mixing process and moments, both of which are commonly employed in the literature (Fan and Yao, 2003). Condition (C.3) consists of regularity conditions for the central limit theorem for dependent processes, with (C.3)(i) ensuring the variance does not degenerate to zero as* $(T-h)l \to \infty$*, and (ii) providing a bound on the moments (see White (2014)). Condition (C.4) requires the two-sided kernel to be symmetric and bounded with a compact support* $[-1,1]$*, contrasting the one-sided kernel with a compact support* $[-1,0]$ *in time-varying coefficient regression models for out-of-sample forecasting. If* $\nu = 1/5$*, the optimal bandwidth* $h_{opt} = O(T^{-1/5})$ *satisfies Condition (C.5).*

Remark 3. *Condition (C.6) is a mild technical requirement, equivalent to Condition 7 in Hu and Zhang (2023) for time-invariant model averaging settings. More discussions and verifications of Condition (C.6) are available in Appendix A. Condition (C.7) bounds the number of candidate models and the number of predictors relative to the sample size, similar to Condition 7 in Ando and Li (2014). Note that it implies* $\xi^*_{T+h}(U_T) > 0$*, requiring all candidate models are misspecified. If the* m_0*-th candidate model is correctly specified, then for any given* u_0 $\boldsymbol{\alpha}^{(m_0)*}(u_0) = \boldsymbol{\alpha}(u_0)$*. Thus,* $\xi^*_{T+h}(U_T) = \inf_{\mathbf{w}\in\mathcal{H}} R^*_{T+h}(\mathbf{w}) \leq \mathbb{E}\left\{Y_{T+h} - Y^{(m_0)*}_{T+h}\right\}^2 - \sigma^2_{T+h} = 0$*, violating Condition (C.7).*

We first discuss asymptotic optimality when all candidate models are misspecified, and then, discuss the alternative cases where some models are correctly specified.

Theorem 1. *Suppose that Conditions (C.1)-(C.7) hold. Then, the SVMA estimator satisfies the asymptotic optimality property, i.e.,*

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

where $\xrightarrow{p}$ *denotes the convergence in probability as* $T \to \infty$*.*

Theorem 1 demonstrates the proposed SVMA prediction for Y_{T+h} achieves asymptotic optimality in the sense of attaining the minimum attainable prediction risk, when all candidate models are misspecified. This underscores the advantages of the proposed method over other averaging or selection methods, including smoothed AIC (SAIC) or smoothed BIC (SBIC), as the infeasible prediction risk of the possibly best model averaging estimator is no larger than that of other model averaging estimators and model selection estimators. Moreover, unlike some previous model averaging literature based on in-sample loss (Hansen, 2007; Wan et al., 2010; Racine et al., 2023; Sun et al., 2021, 2023; Chen et al., 2024), the result of asymptotic optimality is established based on the out-of-sample prediction risk, which is more pertinent to forecasting applications. Departing from the commonly employed Whittle's inequality, we provide a new technique to bound the discrepancy between the state-dependent forward-validation criterion and the prediction risk function.

Condition (C.8). $T^{-1/2}l^{-1/2}(qM + M^2 + q^2MT^{-1/2}l^{-1/2})\{\inf_{\mathbf{w}\in\widetilde{\mathcal{H}}} R^*_{T+h}(\mathbf{w})\}^{-1} = o(1)$, *and* $qMT^{-1}l^{-1} = o(1)$, *where* $\widetilde{\mathcal{H}} = \{\mathbf{w} \in [0,1]^M : \sum_{m\notin\mathcal{D}} w^m = 1\}$ *and* $\mathcal{D}$ *is the subset of* $\{1,\ldots,M\}$, *composed of the correctly specified models.*

Condition (C.9). $\{\inf_{\mathbf{w}\in\widetilde{\mathcal{H}}} R^*_{T+h}(\mathbf{w})\}^{-1} = O(1)$.

Remark 4. *Condition (C.8) constrains the growth rate of the minimum out-of-sample prediction risk, when formulating the averaging prediction by taking the average across all misspecified models. Indeed, Condition (C.8) is essentially equivalent to Condition (C.7), if* $\mathcal{D}$ *is empty. Condition (C.9) also characterizes the quality of the misspecified candidate models, which is a special case of Condition (C.8) given suitable* q *and* M.

Theorem 2. *If there is one or more correctly specified models, and Conditions (C.1)-(C.5) and (C.8) are satisfied, then* $\sum_{m\in\mathcal{D}} \widehat{w}^m(U_T) \xrightarrow{p} 1$.

Theorem 2 implies that our SVMA asymptotically assigns all weights to the correctly specified models when they are included in candidate models. In the case where the model set contains a single correct model, Theorem 2 ensures that the proposed state-dependent forward-validation criterion should asymptotically identify this correct model specification.

Remark 5. *Note that Theorem 1 remains valid for DGPs with discrete state spaces, provided that all candidate models are misspecified and parameters are estimated using local constant estimation. By accommodating discrete state spaces, it can encompass a broader range of applications in economics and related fields, such as business cycles. When the candidate model incorporates a discrete state space, Theorem 2 continues to hold. For each candidate model, we could employ the local constant method to estimate the varying-coefficients conditional on a discrete state outcome. For instance, if we aim to estimate a varying-coefficient candidate model during recessions, utilizing a uniform kernel with a small bandwidth is equivalent to using the data only in recessions to estimate parameters, which is theoretically justified by Lemma 3 in Appendix A based on some slight modifications.*

In data-rich environments with numerous candidate models incorporating diverse predictor sets, predictions are commonly combined via SA (Stock and Watson, 2004). When minimizing mean squared forecast error, a bias-variance tradeoff emerges. Estimated optimal weights may exhibit high variance, potentially underperforming SA (Hsiao and Wan, 2014). However, if one terrible model is included in the candidate model set, the prediction error of the SA can deteriorate rapidly (Hansen, 2007; Qian and Raphael, 2019). Notably, we establish that our method dominates the SA, as formalized in the following Corollary 1.

Corollary 1. *Assume* $T^{-1}l^{-1}M^3q^2K^{-2} = o(1)$ *and* $T^{-1/2}l^{-1/2}K^{-2}(qM^{5/2} + M^4) = o(1)$, *where* K *is the number of misspecified models. Under Conditions (C.1)-(C.7) and (C.9),*

regardless of whether the correctly specified models are included in the candidate set, the prediction risk of the SVMA is smaller than the risk associated with the SA.

If the number of candidate models M increases, it is plausible to assume that the number of misspecified models K also increases. Corollary 1 provides a theoretical justification for the superior performance of dynamic weights compared to SA, offering a novel perspective to resolve the forecast combination puzzle. This rationale has remained unexplored in the existing literature on time-varying model averaging (Sun et al., 2023; Chen et al., 2024).

3 Extensions

3.1 SVMA for Ultra-High Dimensional Setting

In this subsection, we consider the case that both the number of predictors and the number of candidate models are allowed to grow to infinity at some slower rates than the sample size T. If the number of candidate models M is large (e.g., $M > T$), the computational burden of our state-varying model averaging procedure should be heavy. This section is devoted to addressing the dimensionality issue encountered in regression problems with $q > T$ and reduce the computational burden of model averaging procedure.

Our procedure consists of two steps. In Step 1, we use model screening to prepare candidate models, which essentially selects a valid subset of all candidate models. Various model screening strategies are proposed in the existing literature, including top s model screening in Yuan and and (2005) and ordering model screening in Claeskens et al. (2006). For example, we explore the proposal of Ando and Li (2014, 2017) for model screening as a special case, which calculates the marginal correlation between each predictor and the

dependent variable, without prior subject knowledge or expert theories. In Step 2, we construct model averaging based on the subset M^*, where M^* is a subset of $\{1,\ldots,M\}$, and the weight vector is derived from $\widehat{\mathbf{w}}^*(u_0) = \arg\min_{\mathbf{w}\in\mathcal{H}^*} \mathrm{SFV}(u_0,\mathbf{w})$, where $\mathcal{H}^* = \{\mathbf{w} \in [0,1]^{M^*} : \sum_{m\in M^*} w^m = 1 \text{ and } \sum_{m\notin M^*} w^m = 0\}$, which will be shown to be asymptotically optimal under some regularity conditions, presented in Corollary 2 later.

Condition (C.10). *There exist a nonnegative series of $v_T(U_T)$ and a weight series of $\mathbf{w}_T(U_T) \in \mathcal{H}$ such that $\xi_{T+h}^{*-1}(U_T)v_T(U_T) \to 0$, $\inf_{\mathbf{w}\in\mathcal{H}} R_{T+h}(\mathbf{w}) = R_{T+h}(\mathbf{w}_T(U_T)) - v_T(U_T)$, and $\Pr(\mathbf{w}_T(U_T) \in \mathcal{H}^*) \to 1$ as $T \to \infty$.*

Condition (C.10) requires that the infeasible time-variant optimal weight vector asymptotically belongs to the screened set $\mathcal{H}^*$ with probability approaching to 1. This requirement bears resemblance to Assumption 0 of Zhang et al. (2016) and Lemma 3 of Liao et al. (2021) for time-invariant combination weights.

Corollary 2. *Suppose Conditions (C.1)-(C.7) and (C.10) hold. Then, we have*

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

Corollary 2 establishes that, under Condition (C.10) and other regularity conditions, the SVMA estimator in this subsection attains asymptotic optimality in ultra-high dimensional settings when the state-dependent combination weights are selected from the candidate model set $\mathcal{H}^*$. Specifically, the prediction risk of our SVMA estimator converges to that of the infeasible best model averaging, which selects weights from the ex-ante model set $\mathcal{H}$.

Remark 6. *When the number of parameters q_m diverges as T increases, alternative penalized estimation methods can be employed for the candidate models. For instance, one can adopt the approach of Lehrer and Xie (2017) and Kapetanios and Zikes (2018) and utilize*

time-invariant or time-variant LASSO-type penalized estimation techniques. Furthermore, we conjecture that the asymptotic optimality established in Theorem 1 can be extended to encompass such penalized estimation procedures given regularity conditions.

3.2 SVMA for State-Varying Factor Models

This subsection discusses the case where all candidate models are varying-coefficient factor-augmented models, allowing for state-varying factor loadings to capture state-dependency. Let $\{\mathbf{V}_{it}, i = 1, \ldots, N, t = 1, \ldots, T\}$ be an N-dimensional time series with T observations in addition to the observations $\{\mathbf{X}_t, U_t, Y_t\}_{t=1}^{T}$. We forecast Y_{t+h} using the varying-coefficient factor-augmented model $Y_{t+h} = \mathbf{X}_t'\boldsymbol{\alpha}(\mathbf{U}_t) + \mathbf{F}_t'\boldsymbol{\beta}(\mathbf{U}_t) + \epsilon_{t+h}$, where $\mathbf{F}_t$ is the r-dimensional vector of latent common factors. Following Pelger and Xiong (2022), $\mathbf{F}_t$ is estimated by the state-varying factor model $\mathbf{Z}_{it} = \boldsymbol{\lambda}_i(U_t)'\mathbf{F}_t + \mathbf{e}_{it}$, where $\mathbf{Z}_{it}$ is a $d_z \times 1$ vector of observable information, which might contain $\mathbf{X}_t$ and $\mathbf{V}_{it}$, $\boldsymbol{\lambda}_i(U_t)$ is the $r \times d_z$ matrix of factor loadings of the cross-sectional unit i when the state value is U_t, and $\mathbf{e}_{it}$ is idiosyncratic component.

Next, we follow the same idea as in Pelger and Xiong (2022) to estimate factor loadings and factors in state-varying factor models, and then employ local constant estimation to estimate the varying-coefficient models as in Section 2. We impose no restrictions on the approximating models; specifically, the models may be nested or non-nested, and they may incorporate all r factors, only a subset, or even none. For simplicity, we consider the case where all factors are included with different lags[3]. That is, the m-th candidate model is $Y_{t+h} = \sum_{j=0}^{q_m-1} \alpha_j^{(m)}(U_t)X_{tj} + \sum_{j=0}^{r_m-1} \boldsymbol{\beta}_j^{(m)'}(U_t)\mathbf{F}_{t-j} + \epsilon_{t+h}^{(m)}$. Since $\mathbf{F}_t$ is unobservable, we replace $\mathbf{F}_t$ with a generic estimator $\check{\mathbf{F}}_t$, which is consistent up to the state-dependent rotation matrix

[3]Note that the theoretical proof of the general case is similar to that for this simpler case.

$\mathbf{H}(u_0)$ given any u_0. Then, the m-th candidate model implies

$$Y_{t+h} = \sum_{j=0}^{q_m-1} \alpha_j^{(m)}(U_t)X_{tj} + \sum_{j=0}^{r_m-1} \boldsymbol{\beta}_j^{(m)'}(U_t)\mathbf{H}^{'-1}(U_t)\check{\mathbf{F}}_{t-j} + \epsilon_{t+h}^{(m)} = \mathbb{Z}_t^{(m)'}\boldsymbol{\Psi}(U_t) + \epsilon_{t+h}^{(m)},$$

where $\mathbb{Z}_t^{(m)} = (X_{t0}, \ldots, X_{t(q_m-1)}, \check{\mathbf{F}}_t', \ldots, \check{\mathbf{F}}_{t-r_m+1}')'$ and $\boldsymbol{\Psi}(U_t) = (\alpha_0^{(m)}(U_t), \ldots, \alpha_{q_m-1}^{(m)}(U_t),$ $\boldsymbol{\beta}_0^{(m)'}(U_t)\mathbf{H}^{'-1}(U_t), \ldots, \boldsymbol{\beta}_{r_m-1}^{(m)'}(U_t)\mathbf{H}^{'-1}(U_t))'$.

For each candidate model, we use the following two-stage estimation. In the first stage, the local principal component analysis (PCA) approach proposed by Pelger and Xiong (2022) is used to estimate state-varying factor model. For simplicity of notation, assume that $d_z = 1$ and then $\mathbf{Z}_{it}$ is changed to Z_{it}. Then, for any given u_0, let the $(T-h) \times N$ matrix $\mathbf{Z}(u_0) = (\mathbf{Z}_1(u_0), \ldots, \mathbf{Z}_N(u_0))$ with $\mathbf{Z}_i(u_0) \equiv (k_{h+1}^{1/2} Z_{i(h+1)}, \ldots, k_T^{1/2} Z_{iT})'$, the $(T-h)\times r$ matrix $\mathbf{F}(u_0) = (k_{h+1}^{1/2}\mathbf{F}_{h+1}, \ldots, k_T^{1/2}\mathbf{F}_T)'$, and the $N \times r$ matrix $\boldsymbol{\Lambda}(u_0) = (\boldsymbol{\lambda}_1(u_0), \ldots, \boldsymbol{\lambda}_N(u_0))'$. The state-varying factor loadings and the common factors can be consistently estimated by solving the following minimization problem

$$\min_{\mathbf{F}(u_0), \boldsymbol{\Lambda}(u_0)} \operatorname{tr}\left([\mathbf{Z}(u_0) - \mathbf{F}(u_0)\boldsymbol{\Lambda}'(u_0)] [\mathbf{Z}(u_0) - \mathbf{F}(u_0)\boldsymbol{\Lambda}'(u_0)]' \right),$$

subject to $(T-h)^{-1}l^{-1}\mathbf{F}(u_0)'\mathbf{F}(u_0) = \mathbf{I}_r$. The eigenvectors corresponding to the r largest eigenvalues of $\mathbf{Z}(u_0)\mathbf{Z}'(u_0)$ form the estimated factor matrix $\widehat{\mathbf{F}}(u_0)$. Note that the estimated common factor $\widehat{\mathbf{F}}(u_0)$ is a consistent estimator for the latent factor $\mathbf{F}_t$ up to state-dependent rotation matrix $\mathbf{H}(u_0)$ for any given u_0; see Pelger and Xiong (2022), Su and Wang (2017), and Fu et al. (2024) for more discussions.

In the second stage, for each candidate model $m = 1, \ldots, M$, we employ the local constant estimation method described in Section 2 to obtain the following estimator

$$\widehat{\boldsymbol{\Psi}}^{(m)}(u_0) = \left(\sum_{t=1}^{T-h} k_t \widehat{\mathbb{Z}}_t^{(m)}(u_0)\widehat{\mathbb{Z}}_t^{(m)'}(u_0) \right)^{-1} \sum_{t=1}^{T-h} k_t \widehat{\mathbb{Z}}_t^{(m)}(u_0) Y_{t+h},$$

where $\widehat{\mathbb{Z}}_t^{(m)}(u_0) = (X_{t0}, \ldots, X_{t(q_m-1)}, \widehat{\mathbf{F}}_t'(u_0), \ldots, \widehat{\mathbf{F}}_{t-r_m+1}'(u_0))'$. Then, the forecast at $T+h$ in the m-th candidate model is $Y_{T+h}^{(m)} = \widehat{\mathbb{Z}}_T^{(m)'} \widehat{\mathbf{\Psi}}^{(m)}(U_T)$. For any given u_0, we minimize $\mathrm{SFV}(u_0, \mathbf{w})$ within $\widehat{\mathbb{Z}}_t^{(m)}$ and $\widehat{\mathbf{\Psi}}_{[-t]}^{(m)}(u_0)$, the parameter estimator after removing observations, to obtain $\widehat{\mathbf{w}}(u_0)$. After that, the model averaging prediction at $T+h$ is $\widehat{Y}_{T+h}(\widehat{\mathbf{w}}(U_T)) = \sum_{m=1}^{M} \widehat{w}^m(U_T) \widehat{Y}_{T+h}^{(m)}$. See more detailed discussions in Appendix A.

Next, we discuss the asymptotic optimality of the SVMA estimator for varying-coefficient factor-augmented regressions. All convergences occur as $(T-h)l$ and N go to infinity and we assume both $r_0 = \max\{r_1, \ldots, r_M\}$ and r are fixed and thus $O(q + r \times r_0) = O(q)$.

Corollary 3. *Suppose Conditions (C.4)-(C.7) and (C.11)-(C.12) in Appendix B.3 hold. Then, the SVMA estimator for the varying-coefficient factor-augmented case still satisfies the asymptotic optimality property.*

Corollary 3 extends the asymptotic optimality of the SVMA estimator, established in Theorem 1, to varying-coefficient factor-augmented regressions with state-dependent factor dynamics. Unlike existing model averaging for factor-augmented regressions with time-invariant (Cheng and Hansen, 2015) and time-variant weights (Chen et al., 2024), we allow the loadings to be general functions of a recurrent state process, rather than time-invariant. Importantly, the asymptotic optimality is established based on the prediction risk contrasting the in-sample loss criteria employed in Cheng and Hansen (2015) and Chen et al. (2024).

4 Simulation Studies

This section evaluates the finite-sample performance of the proposed SVMA method against competing methods for horizons $h = 1$, 2 and 4, including the SA with equal weighting for each model, the nonparametric version of bias-corrected AIC (AICc) model selection

by Cai and Tiwari (2000), the smoothed AICc (SAICc), the SAIC and the SBIC.

Our simulation study is based on the following DGP framework:

$$Y_{t+h} = \sum_{j=1}^{q} \alpha_j(X_{t1}) X_{tj} + \epsilon_{t+h}, \quad t = 1, \ldots, T, \tag{4}$$

where $\alpha_j(\cdot)$, X_{tj} and ϵ_{t+h} have different specifications given in Examples 1-4. For simplicity, we consider a set of nested candidate models and omit the largest model. Consequently, the number of candidate models is $q-1$, and all of these models are misspecified.

Example 1: We consider model (4) with $\alpha_j(u) = [1+\exp(-cu/j)]^{-1}$ and $\epsilon_t \sim i.i.d.N(0, 0.3^2)$. Predictors follow ARMA processes: $X_{t1} = 0.8X_{(t-1)1} + v_{t1}$, $v_{t1} \sim i.i.d.N(0,1)$ and $X_{t2} = X_{(t-1)1}$. Furthermore, $X_{t3} = 0.6X_{(t-1)3} + 0.3v_{(t-1)3} + v_{t3}$, $v_{t3} \sim i.i.d.N(0,1)$, $X_{t4} = X_{(t-1)3}$, and $X_{t5} = X_{(t-2)3}$. Thus, some predictors are lagged variables. Additionally, for $j > 5$, $X_{tj} = (-0.3 + 0.1j)X_{(t-1)j} + v_{tj}$, where $v_{tj} \sim i.i.d.N(0, s_j^2)$ and s_j is a random sample generated from the Chi-square distribution χ_1^2. We set $c = 2$ and the number of predictors $q = 10$. The proposed DGP in this example includes varying-coefficients $\alpha_j(\cdot)$ that depend on the index j and are modeled as *logit* functions. This design allows predictor contributions to the conditional mean to vary. Specifically, the varying form of $\alpha_j(\cdot)$ is such that the curvature of the *logit* function tends to flatten as j increases. Consequently, this setting encompasses both linear and nonlinear types of varying-coefficients.

Example 2: The setup for this simulation example is adapted from Example 1, but with varying-coefficients $\alpha_j(u) = [1+\exp((-1)^j \cdot cu/j)]^{-1}$ for $j = 1, \ldots, q$, where odd j coefficients are decreasing and even j coefficients are increasing. This approach achieves the conversion of monotonicity of the coefficients through the parity of j.

Example 3: Our setting is similar to Example 1, except for the varying-coefficients $\alpha_j(u)$ and the corresponding predictors. Specifically, for $j = 1,\ 2$, $\alpha_j(u) = \sqrt{2c}j^{-c}\exp(-3u^2)$ with

$X_{t1} = 0.8X_{(t-1)1} + v_{t1}$ and $X_{t2} = X_{(t-1)1}$, where $\alpha_j(u)$ combines a decaying exponential and power function of j. For $j = 3,\ 4,\ 5$, $\alpha_j(u) = (1-\alpha)\alpha^j u$ with $X_{t3} = 0.6X_{(t-1)3} + 0.3v_{(t-1)3} + v_{t3}$, $X_{t4} = X_{(t-1)3}$, and $X_{t5} = X_{(t-2)3}$. For $j > 5$, $\alpha_j(u) = (u^2 - ju)/3j$ adopting a quadratic form with $X_{tj} = (-0.3 + 0.1j)X_{(t-1)j} + v_{tj}$. for $j > 5$. Here, $c = 1$, $\alpha = 0.95$, and $q = 8$.

Example 4: We consider a diverging dimensionality scenario where the number of predictors q grows as $q = \lfloor 3T^{1/3} \rceil$ as the sample size $T \to \infty$, where $\lfloor x \rceil$ denotes the rounding of x to the nearest integer. The predictor setting follows Example 1, and the coefficient function $\alpha_j(u) = \sqrt{2}j^{-1}\exp(-3u^2)$ shrinks to zero as j increases, mimicking nuisance predictors weakly associated with the response in high dimensions. Their inclusion may increase variance and reduce efficiency in estimating coefficients of interest. These predictors act as nuisance variables, adding noise without contributing significantly to estimation.

For SVMA, SA, AICc and SAICc methods, the m-th candidate models are specified as follows: $Y_{t+h} = \sum_{j=1}^{m} \alpha_j^{(m)}(X_{t1})X_{tj} + \epsilon_{t+h}$, where the varying-coefficients $\alpha_j^{(m)}(X_{t1})$ are estimated using the local constant approach described in Section 2.2 and $m = 1, \ldots, q-1$. However, these approaches differ in their criteria for selecting the weights assigned to each candidate model. For all simulation experiments, we employ the Epanechnikov kernel function $K(u) = 0.75(1-u^2)\mathbb{I}(|u| \leq 1)$, which assigns diminishing weights to observations lying further from the point of estimation within each localized subsample. The bandwidth parameter is selected via the common rule-of-thumb $l = 2.34S_{X_{t1}}T^{-1/5}$, where $S_{X_{t1}}$ denotes the sample standard deviation of the state variable X_{t1}. Simulations are also conducted using the Gaussian kernel, but the results are similar, thus omitted to save space.

On the other hand, for the SAIC and SBIC methods, the varying-coefficients are assumed to be constant within each candidate model, thereby simplifying the models to

$Y_{t+h} = \sum_{j=1}^{m} \alpha_j^{(m)} X_{tj} + \epsilon_{t+h}$ for $m = 1, \ldots, q-1$. The constant coefficients are estimated via ordinary least squares (OLS), and time-invariant weights are selected for these models. We use 1000 replications and set the sample size $T = 200$ and 600. For each replication, we compute the following out-of-sample forecast criterion MSFE $= \frac{1}{1000} \sum_{k=1}^{1000} \left(\widehat{Y}_{T+h}^{(k)} - Y_{T+h}^{(k)} \right)^2$, where $\widehat{Y}_{T+h}^{(k)}$ is the out-of-sample forecast of $Y_{T+h}^{(k)}$ in the k-th simulation. The numerical results are presented in Table 1 for Examples 1-4. To facilitate comparison, we highlight the best-performing strategy for each case in bold.

From Table 1, it is observed that for all cases, the SVMA method consistently ranks first among the other five competing methods. This outcome can be attributed to the asymptotic optimality of the proposed weight estimator in the SVMA method, which achieves the lowest possible out-of-sample prediction risk, as discussed in Section 2.4. Besides, when one of the candidate models is very close to the correctly specified model, the SVMA method exhibits significantly smaller MSFE compared to other model averaging methods. For instance, in Example 1 with $q = 10$ and $T = 200$, the MSFE of the SVMA approach is 0.3482, substantially lower than the MSFE of 5.2842 obtained by the SA method. Third, the improvement of SVMA over the AICc model selection criterion in Examples 1-2, which exhibits the second-best forecast performance, ranges from 6% to 50%. This underscores the significance of model averaging, which mitigates model uncertainty and serves as a form of insurance against selecting a suboptimal candidate model. Furthermore, the SVMA outperforms the SAICc, with the reduction in MSFE reaching up to tenfold in certain instances. This superior performance of SVMA can be attributed to its state-dependent combination weights, which acknowledges that the forecasting ability of individual candidate models may change over different states of the economy or market conditions. By allowing for dynamic model weights that adapt to the state variables, SVMA effectively captures the potential

Table 1: Simulation results for Examples 1-4

Example 1		SVMA	SA	AICc	SAICc	SAIC	SBIC
h	T						
1	200	**0.3482**	5.2842	0.3846	1.7474	7.0876	7.1159
	600	**0.1469**	5.1484	0.1570	1.2254	6.7646	6.7646
2	200	**0.2885**	4.7695	0.3335	1.5681	6.8454	6.9497
	600	**0.2035**	4.8370	0.2891	1.2744	6.6130	6.6130
4	200	**0.4068**	5.1670	0.5028	1.7285	6.6753	6.7322
	600	**0.1792**	5.4351	0.1905	1.3128	7.1993	7.1993
Example 2		SVMA	SA	AICc	SAICc	SAIC	SBIC
1	200	**0.2045**	4.8070	0.4359	1.4180	1.1613	1.1613
	600	**0.1226**	4.3865	0.1334	0.9911	1.0664	1.0664
2	200	**0.1869**	4.5621	0.2022	1.2952	1.1156	1.1156
	600	**0.1334**	4.6761	0.1411	1.0920	1.1783	1.1783
4	200	**0.1854**	4.7398	0.1917	1.3452	1.0962	1.0962
	600	**0.1196**	4.3674	0.1291	1.0012	1.1308	1.1308
Example 3		SVMA	SA	AICc	SAICc	SAIC	SBIC
1	200	**0.2517**	0.3302	0.2810	0.2950	0.8789	0.8674
	600	**0.1260**	0.2668	0.1278	0.2123	0.8919	0.8953
2	200	**0.1874**	0.2955	0.2053	0.2620	0.8006	0.8041
	600	**0.1399**	0.2896	0.1897	0.2368	0.9231	0.9266
4	200	**0.1959**	0.3023	0.2127	0.2600	0.8447	0.8326
	600	**0.1437**	0.2788	0.1643	0.2253	0.9044	0.9066
Example 4		SVMA	SA	AICc	SAICc	SAIC	SBIC
1	200	**0.4421**	0.6042	0.6150	0.5607	2.2984	2.3015
	600	**0.2914**	0.4267	0.3680	0.3806	2.6062	2.6497
2	200	**0.4889**	0.6671	0.6233	0.6135	2.5668	2.5265
	600	**0.2876**	0.9009	0.3644	1.0072	2.6924	2.7310
4	200	**0.4694**	0.6204	0.5982	0.5770	2.3763	2.3291
	600	**0.2825**	0.4162	0.3500	0.3578	2.8447	2.8728

shifts, thereby enhancing the forecast accuracy.

5 Empirical Examples

In empirical applications, we compare the performance of the SVMA with competing methods, including AICc, SAICc, SA, SAIC, and SBIC. The candidate model specifications align with our simulation setup. Specifically, for SVMA, AICc, SAICc, and SA, we employ the same set of candidate models, with the parameters estimated using the local constant approach. In contrast, for SAIC and SBIC, the parameters in each candidate model are estimated via OLS, and time-invariant weights are assigned to these models. For simplicity, all candidate models are nested and the number of candidate models is $M = \min(\lfloor 3T^{1/3} \rceil, q)$, where T is the sample size and q is the number of predictors in the full model.

5.1 Stock Return Forecasting

In this subsection, we apply our SVMA method to predict the stock return from monetary policy, compared with other competing methods. The dataset is identical to that used by Jansen et al. (2008) and Tu and Wang (2020), comprising monthly observations of the S&P 500 index, federal funds rate (FF), industrial production (IP), and the US government budget deficit (or surplus, denoted as FD) obtained from the Federal Reserve Economic Data (FRED). The data span from October 1980 to December 2020, with a total of 482 observations. The following variables are utilized based on the transformation of the original data, i.e., stock return (SR_t), the growth in industrial production ($\mathrm{IPG}_t = \ln(\mathrm{IP}_t) - \ln(\mathrm{IP}_{t-1})$), the first difference of the effective federal fund rate ($\mathrm{DFF}_t = \mathrm{FF}_t - \mathrm{FF}_{t-1}$) and the change in fiscal deficits ($\mathrm{CFD}_t = \mathrm{FD}_t - \mathrm{FD}_{t-1}$), where deficits are positive and surpluses are negative.

Following Jansen et al. (2008), we consider the m-th candidate model as follows

$$\mathrm{SR}_{t+h} = \mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(\mathrm{CFD}_t) + \epsilon_{t+h},$$

where $\mathbf{X}_t^{(m)} = (X_{t1}, \ldots, X_{tq_m})'$ is a $q_m \times 1$ vector of regressors. We construct the candidate pool Ω, where $\Omega = \{\mathrm{SR}_t, \ldots, \mathrm{SR}_{t-k_1}, \mathrm{IPG}_t, \ldots, \mathrm{IPG}_{t-k_1}, \mathrm{DFF}_t, \ldots, \mathrm{DFF}_{t-k_1}\}$, $X_{ti} \in \Omega$ and the lag order $k_1 = 5$, which is consistent with the maximum lag order employed in prior studies (Jansen et al., 2008; Tu and Wang, 2020). Thus, $q = 18$ and the number of candidate models $M = 18$. We utilize the Epanechnikov kernel function, and the bandwidth is selected via $l = 2.34 S_{\mathrm{CFD}} T^{-1/5}$, where S_{CFD} is the sample standard deviation of $\{\mathrm{CFD}_t\}$.

Forecasting performance is evaluated via the following relative MSFE gain:

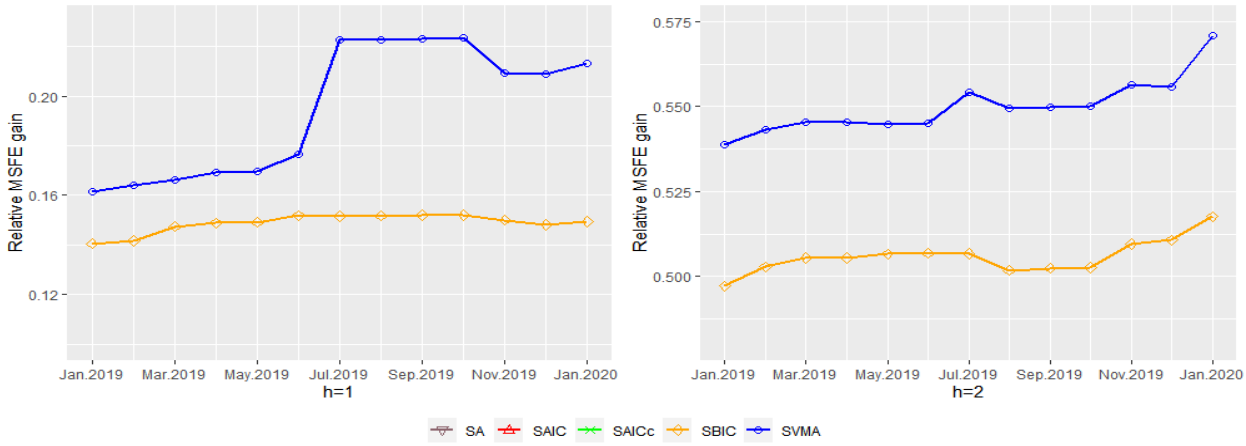
$$\mathrm{MSFE}_{(i)}^{h}(T_1, T_2) = 1 - \frac{\sum_{t=T_1}^{T_2} \widehat{\epsilon}_{t,(i)}^2}{\sum_{t=T_1}^{T_2} \widehat{\epsilon}_{t,(\mathrm{AIC}_c)}^2} \quad \text{with} \quad T_2 = 2020:12,$$

where the forecast data $T_1 \in [2019:1, 2020:1]$, and AICc serves as the benchmark method. Relative percentage gains in MSFE for the method i versus AICc are plotted, with the initial point spanning from 2019:1 to 2020:12 and the final point for 2020:1-2020:12. A positive value implies that the method i produces more accurate forecast than the AICc.

Figure 1 presents the forecast gains achieved by various approaches relative to the AICc method. A notable observation is that the proposed SVMA consistently outperforms its competitors for all cases, yielding improvements ranging from 0.16 to 0.22 for one-step-ahead forecasts and 0.53 to 0.57 for two-step-ahead forecasts over the AICc method. This highlights the merits of reducing model uncertainty and increasing the robustness of model averaging. Interestingly, the OLS-based SAIC and SBIC approaches exhibit a better performance compared to the nonparametric-based SAICc, SA, and AICc methods. A plausible explanation for this phenomenon is that nonparametric techniques may induce excessive

variance, thereby compromising the accuracy of out-of-sample forecasts. More importantly, the proposed SVMA outperforms SAIC and SBIC, attributable to the asymptotic optimality of the SVMA prediction achieving the lowest possible out-of-sample prediction risk among all model averaging estimators. Furthermore, the SVMA significantly outperforms the SA, despite both approaches facing estimation variance from nonparametric candidate models. This happens because the SVMA effectively mitigates sufficient bias through optimal weight assignment, consequently enhancing forecast accuracy, which confirms Corollary 1.

Figure 1: Relative MSFE gains: the left panel for $h = 1$ and the right panel for $h = 2$.



Note: Results exceeding the vertical axis limits in Figure 1, such as the negative forecast gains of SAIC and SAICc, are omitted but reported in Table C.1 in Appendix C.

5.2 Revisiting Macroeconomic Forecasting

In this subsection, our purpose is to revisit the so-called forecast combination puzzle as discussed in Stock and Watson (2004), where simple constant averaging outperforms sophisticated adaptive weighting. To this end, we implement the proposed methodology to forecast macroeconomic indicators using the FRED-MD database, as described in McCracken and Ng (2016). Series with missing values were removed, resulting in a refined dataset of

119 monthly macroeconomic variables. For robustness, our analysis focuses on three key macroeconomic variables: Real Personal Income (RPI), 3-Month Treasury Bill (TB3MS) and Total Non-farm Payroll (PAYEMS). Additionally, we incorporate the monthly economic policy uncertainty (EPU) index for the United States[4] as the state-variable in our analysis. The monthly data spans from January 1985 to December 2023, aligned with the availability of the EPU index.

The formulation of the target variable and forecasting variables follows Yousuf and Ng (2021). Specifically, let Y_{t+h} denote the h-step-ahead target variable to be forecasted. For instance, for RPI, our target variable is $Y_{t+h} = \frac{1200}{h}\left[\ln(\mathrm{RPI}_{t+h}) - \ln(\mathrm{RPI}_t)\right]$, and we define the target similarity for PAYEMS, whereas TB3MS is modeled as $I(1)$ (i.e., $Y_{t+h} = \frac{12}{h}\left[\mathrm{TB3MS}_{t+h} - \mathrm{TB3MS}_t\right]$). Let $\mathbf{z}_t$ represent the remainder of our 118 predictor series at time t. The set of predicting variables is denoted as $\mathbf{X}_t = (Y_t, \ldots, Y_{t-3}, \mathbf{F}_t', \ldots, \mathbf{F}_{t-3}')$, where the factor vector $\mathbf{F}_t = (F_{1t}, \ldots, F_{4t})'$. $\mathbf{F}_t$ is estimated by using the principal components of our predictor series $\mathbf{z}_t$. Then, we consider the candidate model

$$\mathrm{Y}_{t+h} = \mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(\mathrm{EPU}_t) + \epsilon_t,$$

where EPU_t represents the economic policy uncertainty index at time t, utilized as a conditioning variable that can dynamically influence the coefficients $\boldsymbol{\alpha}^{(m)}(\cdot)$ and $h = 1$. The candidate models are indexed by m with $m = 1, 2, \ldots, M$, and $\mathbf{X}_t^{(m)} \subset \mathbf{X}_t$ denotes the subset of predictors included in the m-th candidate model. In the current study, this configuration yields $M = 20$, with the largest candidate model containing $q = 20$ predictors. The setup described previously is consistently applied across all target variables.

To evaluate the forecasting performance of our models, we employ the relative MSFE

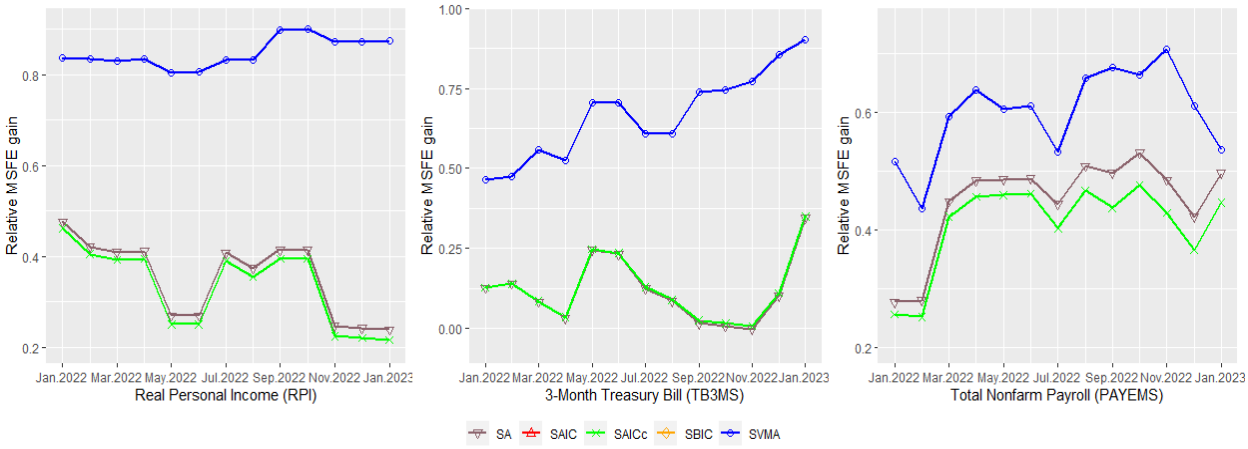
[4]https://fred.stlouisfed.org/series/USEPUINDXM

gains as defined in Section 5.1. Specifically, we calculate the percentage gains by comparing the MSFE of the forecasts produced by each method to those generated using the AICc benchmark. In this application, the forecasting performance evaluation is set within a defined timeframe, where T_1 spans from 2022:1 to 2023:1, and T_2 is set to 2023:12.

Figure 2 presents the forecast gains over the AICc benchmark. First, it is observed that the SVMA yields the largest relative MSFE gains for all cases. For RPI, TB3MS and PAYEMS, its gain ranges from 0.80 to 0.90, 0.46 to 0.90 and 0.43 to 0.70 respectively, exceeding all competitors. Second, the SVMA method has better prediction accuracy than the SAICc method. The key distinction between SAICc and SVMA lies in their weight selection criteria: SAICc employs time-invariant weights, while SVMA assigns state-dependent weights. This result underscores the importance of allowing weights to vary across states for each candidate model. A potential explanation is that candidate models are likely misspecified, with their predictive ability rankings fluctuating across different states. The proposed SVMA assigns higher weights to candidate models that perform well in a given state, while attributing lower or even zero weights when their performance is poor in other states. Consequently, the SVMA permits weights to adapt to the state variable, thereby constructing asymptotic optimality as established in Theorem 1. Furthermore, SA performs similarly to SAICc but worse than SVMA, contradicting the forecast combination puzzle from Stock and Watson (2004), where simple averaging outperforms sophisticated adaptive weighting. One possible explanation is that SA reduces variance but increases bias when including poor models, degrading performance, which is consistent with Hansen (2008). Conversely, SVMA increases variance but optimally weights models to reduce bias, potentially improving forecasts over SA if bias reduction outweighs variance increase.

Finally, to gauge the advantages of the proposed method, the Diebold-Mariano (DM)

Figure 2: Relative MSFE gains for macroeconomic indicators



Note: Negative results exceeding the vertical axis limits are omitted from Figure 2 but documented in Table C.2 in Appendix C.

test is implemented to check the statistical significance of differences in forecast accuracy between SVMA and its five competing methods from January 2022 to December 2023. For each pairwise comparison, the test posits the null hypothesis of equal expected forecast error loss, against the one-sided alternative hypothesis that SVMA generates statistically superior accuracy. Table 2 reports p-values for each test. We observe that SVMA demonstrates universal dominance, achieving statistically significant superiority over all competing methods at the 1% or 5% significance levels in most cases.

Table 2: p-values of the DM test

	SA	AICc	SAICc	SAIC	SBIC
RPI	0.0003	0.0002	0.0003	0.0070	0.0271
TB3MS	0.0009	0.0165	0.0009	0.0012	0.0009
PAYEMS	0.0627	0.0126	0.0334	0.0076	0.0090

6 Conclusion

This paper proposes a novel state-dependent model averaging method for high-dimensional varying-coefficient regression models, which allows the selected weights to change over state variables. We establish the asymptotic optimality of the proposed estimator when all candidate models with high-dimensional covariates are misspecified. When the model set includes the correctly specified models, the proposed method asymptotically assigns all weights to the correctly specified models. Also, model screening prior to model averaging in ultra-high-dimensional regressions and averaging in factor-augmented regression models have been investigated. Numerical analyses and empirical applications strongly favor the proposed model averaging in comparison with the existing conventional methods.

Some relevant issues deserve further research. First, it would be interesting to study local optimal averaging for varying-coefficient model to mitigate model uncertainty arising from diverse state variables, beyond just high-dimensional covariates. For example, one can follow Cai et al. (2015) to select state variables prior to model averaging. Besides, one possible extension is to generalize our SVMA approach for nonstationary time series data, which covers more applications in economics and statistics (Xiao, 2009).

References

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association 109*(505), 254–265.

Ando, T. and K.-C. Li (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *Annals of Statistics 45*(6), 2654–2679.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Cai, Z. (2010). Functional coefficient models for economic and financial data. In *Oxford Handbook of Functional Data Analysis* (Eds: F. Ferraty and Y. Romain), Oxford University Press, Oxford, pp. 166 – 186.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*(451), 941–956.

Cai, Z., Y. Ren, and B. Yang (2015). A semiparametric conditional capital asset pricing model. *Journal of Banking & Finance 61*(December), 117–126.

Cai, Z. and R. Tiwari (2000). Application of a local linear autoregressive model to bod time series. *Environmetrics 11*(3), 341–350.

Chen, Q., Y. Hong, and H. Li (2024). Time-varying forecast combination for factor-augmented regressions with smooth structural changes. *Journal of Econometrics 240*(1), 105693.

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

Claeskens, G., C. Croux, and J. van Kerckhoven (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics 62*(4), 972–979.

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

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York.

Fang, F., W. Lan, J. Tong, and J. Shao (2019). Model averaging for prediction with fragmentary data. *Journal of Business & Economic Statistics 37*(3), 517–527.

Fang, F., C. Yuan, and W. Tian (2023). An asymptotic theory for least squares model averaging with nested models. *Econometric Theory 39*(2), 412–441.

Fu, Z., L. Su, and X. Wang (2024). Estimation and inference on time-varying FAVAR models. *Journal of Business & Economic Statistics 42*(2), 533–547.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*(2), 357–384.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*(2), 342–350.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics 5*(3), 495–530.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

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

Hu, X. and X. Zhang (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research 24*(358), 1–53.

Jansen, D. W., Q. Li, Z. Wang, and J. Yang (2008). Fiscal policy and asset markets: A semiparametric analysis. *Journal of Econometrics 147*(1), 141–150.

Jiang, B., J. Lv, J. Li, and M.-Y. Cheng (2024). Robust model averaging prediction of longitudinal response with ultrahigh-dimensional covariates. *Journal of the Royal Statistical Society Series B*, with DOI: https://doi.org/10.1093/jrsssb/qkae094.

Kapetanios, G. and F. Zikes (2018). Time-varying lasso. *Economics Letters 169*, 1–6.

Lehrer, S. and T. Xie (2017). Box office buzz: Does social media data steal the show from model uncertainty when forecasting for hollywood? *Review of Economics and Statistics 99*(5), 749–755.

Li, J., J. Lv, A. T. Wan, and J. Liao (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association 117*(537), 495–509.

Li, J., X. Xia, W. K. Wong, and D. Nott (2018). Varying-coefficient semiparametric model averaging prediction. *Biometrics 74*(4), 1417–1426.

Li, K.-C. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics 15*(3), 958 – 975.

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica* (4), 1563–1588.

Liao, J., G. Zou, Y. Gao, and X. Zhang (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics 223*(1), 190–221.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Paye, B. S. (2012). 'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics 106*(3), 527–546.

Pelger, M. and R. Xiong (2022). State-varying factor models of large dimensions. *Journal of Business & Economic Statistics 40*(3), 1315–1333.

Racine, J. S., Q. Li, D. Yu, and L. Zheng (2023). Optimal model averaging of mixed-data kernel-weighted spline regressions. *Journal of Business & Economic Statistics 41*(4), 1251–1261.

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

Su, L. and X. Wang (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics 198*(1), 84–101.

Sun, Y., Y. Hong, T.-H. Lee, S. Wang, and X. Zhang (2021). Time-varying model averaging. *Journal of Econometrics 222*(2), 974–992.

Sun, Y., Y. Hong, S. Wang, and X. Zhang (2023). Penalized time-varying model averaging. *Journal of Econometrics 235*(2), 1355–1377.

Tu, Y. and Y. Wang (2020). Adaptive estimation of heteroskedastic functional-coefficient regressions with an application to fiscal policy evaluation on asset markets. *Econometric Reviews 39*(3), 299–318.

Wan, A. T. K., X. Zhang, and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics 156*(2), 277–283.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

White, H. (2014). *Asymptotic Theory for Econometricians*. Academic Press, New York.

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics 152*(2), 81–92.

Yousuf, K. and S. Ng (2021). Boosting high dimensional predictive regressions with time varying parameters. *Journal of Econometrics 224*(1), 60–87.

Yuan, Z. and Y. Y. and (2005). Combining linear regression models. *Journal of the American Statistical Association 100*(472), 1202–1214.

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

Zhang, X. and X. Zhang (2023). Optimal model averaging based on forward-validation. *Journal of Econometrics 237*(2), 105295.

Zhu, R., A. T. K. Wan, X. Zhang, and G. Zou (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association 114*(526), 882–892.

Appendix A

A.1 SVMA Algorithm

Algorithm 1: The state-varying model averaging (SVMA) prediction algorithm

Step 1: Calculate the h-step-ahead prediction of Y_{T+h} under each candidate model.

for $m = 1, 2, \ldots, M$

For any given u_0, calculate $\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) = \left(\mathbf{X}^{(m)\prime}\mathbf{K}(u_0)\mathbf{X}^{(m)}\right)^{-1}\mathbf{X}^{(m)\prime}\mathbf{K}(u_0)\mathbf{Y}$; then calculate $\widehat{Y}_{T+h}^{(m)} = \mathbf{X}_T^{(m)\prime}\widehat{\boldsymbol{\alpha}}^{(m)}(U_T)$.

end

Step 2: Calculate the leave-h-out forward-validation estimator for μ_t under every candidate model.

for $m = 1, 2, \ldots, M$

for $t = 1, 2, \ldots, T-h$

For any given u_0, calculate $\widehat{\boldsymbol{\alpha}}_{[-t]}^{(m)}(u_0) = \left(\mathbf{X}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\mathbf{X}_{[-t]}^{(m)}\right)^{-1}\mathbf{X}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\mathbf{Y}_{[-(t+h)]}$. Then, compute $\widetilde{\mu}_t^{(m)}(u_0)$ by Eq. (3).

end

Calculate $\widetilde{\boldsymbol{\mu}}^{(m)}$.

end

Step 3: For any given u_0, solve the constrained quadratic programming problem to obtain the state-dependent weight

$$\widehat{\mathbf{w}}(u_0) = \arg\min_{\mathbf{w}\in\mathcal{H}}\mathrm{SFV}(u_0, \mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{H}}\frac{1}{(T-h)l}(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(u_0)(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})).$$

Step 4: Calculate the SVMA prediction of Y_{T+h}

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

where $\widehat{w}^m(U_T)$ is the m-th element of $\widehat{\mathbf{w}}(U_T)$.

Output: $\widehat{Y}_{T+h}(\widehat{\mathbf{w}}(U_T))$

Note: The *quadprog* package in R language and the *quadprog* command in MATLAB are commonly used to solve such problems.

A.2 Discuss of Condition (C.6)

We discuss whether Condition (C.6) can be satisfied. Define $L(U_T, \mathbf{w}) = \left[Y_{T+h} - \widehat{Y}_{T+h}(\mathbf{w})\right]^2 - \sigma^2_{T+h}$ and $L^*(U_T, \mathbf{w}) = \left[Y_{T+h} - Y^*_{T+h}(\mathbf{w})\right]^2 - \sigma^2_{T+h}$. We have

$$
\begin{aligned}
& \xi^{*-1}_{T+h}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} |L(U_T, \mathbf{w}) - L^*(U_T, \mathbf{w})| \\
= \; & \xi^{*-1}_{T+h}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \left| \left\{ (Y^*_{T+h}(\mathbf{w}) - \widehat{Y}_{T+h}(\mathbf{w}))(2Y_{T+h} - \widehat{Y}_{T+h}(\mathbf{w}) - Y^*_{T+h}(\mathbf{w})) \right\} \right| \\
= \; & \xi^{*-1}_{T+h}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \left| \sum_{i=1}^{M}\sum_{j=1}^{M} w^i w^j \left\{ Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right\} \left\{ 2Y_{T+h} - Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right\} \right| \\
\leq \; & \xi^{*-1}_{T+h}(U_T) \max_{1\leq i\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \max_{1\leq j\leq M} \left| 2Y_{T+h} - Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right| \sup_{\mathbf{w}\in\mathcal{H}} \left\{ \left(\sum_{i=1}^{M} w^i \right) \left(\sum_{j=1}^{M} w^j \right) \right\} \\
= \; & \xi^{*-1}_{T+h}(U_T) \max_{1\leq i\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \max_{1\leq j\leq M} \left| 2\epsilon^{(j)*}_{T+h} + Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right| \\
\leq \; & 2\xi^{*-1}_{T+h}(U_T) \max_{1\leq i,j\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \left| \epsilon^{(j)*}_{T+h} \right| \\
& + \xi^{*-1}_{T+h}(U_T) \max_{1\leq i,j\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \left| Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right| \qquad \text{(A.1)}
\end{aligned}
$$

given that the sum of the combination weights is equal to one.

Condition (C.6) is derived if both $\xi^{*-1}_{T+h}(U_T) \max_{1\leq i,j\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \left| \epsilon^{(j)*}_{T+h} \right|$ and $\xi^{*-1}_{T+h}(U_T)$ $\max_{1\leq i,j\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \left| Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right|$ are uniformly integrable. In the proof of Theorem 1, given Lemma 3, we have that $\xi^{*-1}_{T+h}(U_T) \max_{1\leq m\leq M} \left| \widehat{Y}^{(m)}_{T+h} - Y^{(m)*}_{T+h} \right| = o_p(1)$. Consequently, it is reasonable to deduce the uniform integrability of $\max_{1\leq i,j\leq M} \left| Y^{(i)*}_{T+h} - \widehat{Y}^{(i)}_{T+h} \right| \left| Y^{(j)*}_{T+h} - \widehat{Y}^{(j)}_{T+h} \right|$ under regularity conditions. On the other hand, our proofs establish that $\max_{1\leq m\leq M} \left| \epsilon^{(m)*}_{T+h} \right| = O_p(M^{1/2})$ and $\xi^{*-1}_{T+h}(U_T) \max_{1\leq m\leq M} \left| \widehat{Y}^{(m)}_{T+h} - Y^{(m)*}_{T+h} \right| = O_p(\xi^{*-1}_{T+h}(U_T) M^{1/2} q T^{-1/2} l^{-1/2})$. These results imply that, for certain values of M and q, $\xi^{*-1}_{T+h}(U_T) \max_{1\leq m\leq M} \left| \widehat{Y}^{(m)}_{T+h} - Y^{(m)*}_{T+h} \right| \left| \epsilon^{(m)*}_{T+h} \right| = o_p(1)$ given Condition (C.7). Therefore, Condition (C.6) is satisfied.

A.3 Lemmas 1-3 and Their Proofs

Before we embrace on providing the detailed proof to the main theorems, some lemmas are needed, presented as follows. For simplicity, let $\widetilde{T} = T - h$, where the forecast horizon

h is fixed.

Lemma 1. *Suppose Conditions (C.1)-(C.5) hold. Then, as $T \to \infty$, it holds that*

$$\mathbf{\Psi}_{u_0,\widetilde{T}} \equiv \widetilde{T}^{-1} l^{-1} \sum_{t=1}^{\widetilde{T}} \mathbf{X}_t \mathbf{X}_t' k_t \xrightarrow{p} f_U(u_0) \mathbf{\Psi}(u_0),$$

where $\mathbf{\Psi}(u_0) \equiv \mathbb{E}(\mathbf{X}_t \mathbf{X}_t' | U_t = u_0)$ is a symmetric positive definite matrix.

Proof. This can be directly derived from Theorem 1 in Cai et al. (2000). □

Lemma 2. *Suppose Conditions (C.1)-(C.4) hold. Then for any given u_0, we have*

$$||\widetilde{T}^{-1/2} l^{-1/2} q^{-1/2} \sum_{t=1}^{\widetilde{T}} k_t \mathbf{X}_t \epsilon_{t+h}|| = O_p(1),$$

and

$$||\widetilde{T}^{-1} l^{-1} q^{-1/2} \sum_{t=1}^{\widetilde{T}} k_t \mathbf{X}_t Y_{t+h}|| = O_p(1),$$

if q grows at some slow rate of $\widetilde{T}l$.

Proof. For any given u_0, we have

$$\widetilde{T}^{-1/2} l^{-1/2} q^{-1/2} \sum_{t=1}^{\widetilde{T}} k_t \mathbf{X}_t \epsilon_{t+h} = \widetilde{T}^{-1/2} l^{-1/2} q^{-1/2} \mathbf{X}' \mathbf{K}(u_0) \boldsymbol{\epsilon},$$

and

$$\widetilde{T}^{-1} l^{-1} q^{-1/2} \sum_{t=1}^{\widetilde{T}} k_t \mathbf{X}_t Y_{t+h} = \widetilde{T}^{-1} l^{-1} q^{-1/2} \mathbf{X}' \mathbf{K}(u_0) \mathbf{Y}.$$

Lemma 2 is valid if the following holds:

$$||\mathbf{X}' \mathbf{K}(u_0) \boldsymbol{\epsilon}|| = O_p(\sqrt{q \widetilde{T} l}), \tag{A.2}$$

and

$$||\mathbf{X}' \mathbf{K}(u_0) \mathbf{Y}|| = O_p(\sqrt{q} \widetilde{T} l). \tag{A.3}$$

First, with Conditions (C.1)-(C.4), for any given u_0 we have

$$\begin{aligned}
& \mathbb{E}(\widetilde{T}^{-1}l^{-1}q^{-1}||\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\epsilon}||^2) \\
= & \frac{1}{q\widetilde{T}l}\mathbb{E}(\boldsymbol{\epsilon}'\mathbf{K}(u_0)\mathbf{X}\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\epsilon}) \\
= & \frac{1}{q\widetilde{T}l}\mathrm{tr}\left\{\sqrt{\mathbf{K}(u_0)}\mathbf{X}\mathbf{X}'\sqrt{\mathbf{K}(u_0)}\mathbb{E}(\boldsymbol{\epsilon}'\mathbf{K}(u_0)\boldsymbol{\epsilon})\right\} \\
\leq & \max_{1\leq t\leq\widetilde{T}}\frac{||\mathbf{X}_t||^2}{q}\frac{1}{\widetilde{T}l}\sum_{t=1}^{\widetilde{T}}\mathbb{E}[\mathrm{Var}(Y_{t+h}|U_t=u_0)k_t] \leq C<\infty
\end{aligned}$$

for some positive C and non-stochastic $\mathbf{X}$. We have similar results for random $\mathbf{X}$. Thus, (A.2) holds. Next, with (A.2) and Conditions (C.1)-(C.4), for any given u_0 we have

$$\begin{aligned}
||\mathbf{X}'\mathbf{K}(u_0)\mathbf{Y}|| &= ||\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\mu}+\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\epsilon}|| \\
&\leq ||\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\mu}|| + O_p(\sqrt{q\widetilde{T}l}) \\
&\leq \sum_{t=1}^{\widetilde{T}}|\mu_t|||\mathbf{X}_t||k_t + O_p(\sqrt{q\widetilde{T}l}) \\
&\leq \sqrt{\sum_{t=1}^{\widetilde{T}}\mu_t^2k_t}\sqrt{\sum_{t=1}^{\widetilde{T}}||\mathbf{X}_t||^2k_t} + O_p(\sqrt{q\widetilde{T}l}) \\
&\leq \sqrt{C\widetilde{T}l}\sqrt{\widetilde{T}l\max_{1\leq t\leq\widetilde{T}}||\mathbf{X}_t||^2} + O_p(\sqrt{q\widetilde{T}l}) \\
&= O_p(\sqrt{q}\widetilde{T}l) + O_p(\sqrt{q\widetilde{T}l})
\end{aligned}$$

for some positive constant C. Thus, the proof of (A.3) is completed. □

Lemma 3. *Suppose Conditions (C.1)-(C.5) hold. Then, for any fixed $\varepsilon>0$ and given u_0, there exists a $\delta_\varepsilon>0$ such that for all sufficiently large T,*

$$\Pr\left(\left|\left|\frac{\widetilde{T}^{1/2}l^{1/2}}{q^{1/2}}(\widehat{\boldsymbol{\alpha}}^{(m)}(u_0)-\boldsymbol{\alpha}^{(m)*}(u_0))\right|\right|\leq\delta_\varepsilon\right)\geq 1-\varepsilon.$$

Proof. For the non-stochastic $\mathbf{X}$, we explore the proposal of (A.10) in Zhang et al. (2016) and with Conditions (C.1)-(C.5), we have

$$\Pr\left(\left|\left|\frac{\widetilde{T}^{1/2}l^{1/2}}{q^{1/2}}(\widehat{\boldsymbol{\alpha}}^{(m)}(u_0)-\boldsymbol{\alpha}^{(m)*}(u_0))\right|\right|\leq\delta\right)$$

$$
\begin{aligned}
&\geq \Pr\left(C_0^{-1}\left|\left|\frac{\widetilde{T}^{-1/2}l^{-1/2}}{q^{1/2}}\mathbf{X}^{(m)'}\mathbf{K}(u_0)(\mathbf{Y}-\boldsymbol{\mu})\right|\right| \leq \delta\right) \\
&\geq 1-\frac{\sum_{t=1}^{\widetilde{T}} k_t||\mathbf{X}_t^{(m)}||^2\mathrm{Var}(\epsilon_{t+h})}{C_0^2\delta^2\widetilde{T}lq} \geq 1-\frac{C_1}{C_0^2\delta^2}
\end{aligned}
$$

for some positive constants C_0 and C_1. Thus, the proof of Lemma 3 is completed, when $\delta = \delta_\varepsilon = C_1^{1/2}/(\varepsilon^{1/2}C_0)$.

For the random $\mathbf{X}$, we rewrite $Y_{t+h} = \mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(U_t) + \mathbf{X}_t^{(-m)'}\boldsymbol{\alpha}^{(-m)}(U_t) + \epsilon_{t+h}$, if the m-th model is correctly specified model, then $\boldsymbol{\alpha}^{(-m)}(U_t) = 0$ while if the m-th model is misspecified, then $\mathbb{E}\mathbf{X}_t^{(-m)'}\boldsymbol{\alpha}^{(-m)}(U_t)$ may not be zero. For any given u_0, we have

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) &= [\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{Y} \\
&= [\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(u_0)k_t \\
&\quad +[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\mathbf{X}_t^{(-m)'}\boldsymbol{\alpha}^{(-m)}(u_0)k_t \\
&\quad +[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\epsilon_{t+h}k_t, \qquad \text{(A.4)}
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{\alpha}^{(m)*}(u_0) &= \mathbb{E}[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\mathbb{E}\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{Y} \\
&= \mathbb{E}[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\mathbb{E}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(u_0)k_t \\
&\quad +\mathbb{E}[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\mathbb{E}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\mathbf{X}_t^{(-m)'}\boldsymbol{\alpha}^{(-m)}(u_0)k_t \\
&\quad +\mathbb{E}[\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}]^{-1}\mathbb{E}\sum_{t=1}^{\widetilde{T}}\mathbf{X}_t^{(m)}\epsilon_{t+h}k_t. \qquad \text{(A.5)}
\end{aligned}
$$

Similar to Lemma 1, if q is fixed, then from Lemma 1 and Theorems 1-2 of Cai et al. (2000), we have for any given u_0

$$
\sqrt{\widetilde{T}l}[\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0) - \frac{l^2}{2}\widetilde{c}_2\ddot{\boldsymbol{\alpha}}^{(m)*}(u_0)] \xrightarrow{d} N(\mathbf{0}, \mathbf{V}^{(m)}(u_0)) \qquad \text{(A.6)}
$$

for some positive definite matrix $\mathbf{V}^{(m)}(u_0)$ and some positive $\widetilde{c}_2$, where $\ddot{\boldsymbol{\alpha}}^{(m)*}(u_0)$ is the second derivative of $\boldsymbol{\alpha}^{(m)*}(u_0)$. Subsequently, we discuss if q is diverging along with the sample size T. Based on (A.6), we have for any given u_0

$$\begin{aligned} & \Pr[\sqrt{\widetilde{T}l}||\mathbf{V}^{(m)^{-1/2}}(u_0)(\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0) - \frac{l^2}{2}\widetilde{c}_2\ddot{\boldsymbol{\alpha}}^{(m)*}(u_0))|| \geq \delta q^{1/2}] \\ = \; & \Pr(\chi^2(q_m+1) \geq \delta^2 q) \leq \frac{q_m+1}{\delta^2 q}, \end{aligned} \tag{A.7}$$

and thus $q^{-1/2}\sqrt{\widetilde{T}l}||\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0) - \frac{l^2}{2}\widetilde{c}_2\ddot{\boldsymbol{\alpha}}^{(m)*}(u_0)|| = O_p(1)$. Thus, Lemma 3 is derived.

□

Appendix B

B.1 Proofs of Theorems 1 and 2

Proof of Theorem 1. Let $\mathrm{SFV}^*(U_T, \mathbf{w}) = \mathrm{SFV}(U_T, \mathbf{w}) - \sigma^2_{T+h}$, where σ^2_{T+h} is a constant and unrelated to $\mathbf{w}$. The selected weight

$$\widehat{\mathbf{w}}(U_T) = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{SFV}(U_T, \mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{SFV}^*(U_T, \mathbf{w}). \tag{B.1}$$

From Lemma 1 of Gao et al. (2019), Theorem 1 is valid if

$$\sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{R_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})} - 1 \right| = o(1), \tag{B.2}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}} \left| \frac{\mathrm{SFV}^*(U_T, \mathbf{w}) - R^*_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})} \right| = o_p(1). \tag{B.3}$$

Based on Lemma 3, for any given u_0, it is observed that

$$\max_{1\leq m\leq M} ||\widehat{\boldsymbol{\alpha}}^{(m)}(U_T) - \boldsymbol{\alpha}^{(m)*}(U_T)|| = O_p(M^{1/2}q^{1/2}T^{-1/2}l^{-1/2}).$$

Then, we have

$$\begin{aligned} ||\widehat{\boldsymbol{\alpha}}(U_T,\mathbf{w})-\boldsymbol{\alpha}^*(U_T,\mathbf{w})|| &= \sum_{m=1}^{M} w^m \left|\left|\widehat{\boldsymbol{\alpha}}^{(m)}(U_T)-\boldsymbol{\alpha}^{(m)*}(U_T)\right|\right| \\ &= O_p(M^{1/2}q^{1/2}T^{-1/2}l^{-1/2}). \end{aligned} \tag{B.4}$$

Recall $L(U_T,\mathbf{w})=\left[Y_{T+h}-\widehat{Y}_{T+h}(\mathbf{w})\right]^2-\sigma^2_{T+h}$ and $L^*(U_T,\mathbf{w})=\left[Y_{T+h}-Y^*_{T+h}(\mathbf{w})\right]^2-\sigma^2_{T+h}$ defined in Appendix A.2. Based on Lemma 3, we observe that

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{H}}|L(U_T,\mathbf{w})-L^*(U_T,\mathbf{w})| \\ =\ &\sup_{\mathbf{w}\in\mathcal{H}}\left\{(Y^*_{T+h}(\mathbf{w})-\widehat{Y}_{T+h}(\mathbf{w}))(2Y_{T+h}-\widehat{Y}_{T+h}(\mathbf{w})-Y^*_{T+h}(\mathbf{w}))\right\} \\ =\ &\sup_{\mathbf{w}\in\mathcal{H}}\left\{\mathbf{X}'_T(\boldsymbol{\alpha}^*(U_T,\mathbf{w})-\widehat{\boldsymbol{\alpha}}(U_T,\mathbf{w}))(2Y_{T+h}-\widehat{Y}_{T+h}(\mathbf{w})-Y^*_{T+h}(\mathbf{w}))\right\} \\ =\ &\sup_{\mathbf{w}\in\mathcal{H}}\left\{\mathbf{X}'_T(\boldsymbol{\alpha}^*(U_T,\mathbf{w})-\widehat{\boldsymbol{\alpha}}(U_T,\mathbf{w}))\left[2(Y_{T+h}-Y^*_{T+h}(\mathbf{w}))-(\widehat{Y}_{T+h}(\mathbf{w})-Y^*_{T+h}(\mathbf{w}))\right]\right\} \\ =\ &O_p(MqT^{-1/2}l^{-1/2}+Mq^2T^{-1}l^{-1}), \end{aligned} \tag{B.5}$$

where the last equation is derived from Appendix A.2. Thus,

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{H}}\frac{\left|R_{T+h}(\mathbf{w})-R^*_{T+h}(\mathbf{w})\right|}{R^*_{T+h}(\mathbf{w})} \\ \le\ &\xi^{*-1}_{T+h}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\left|R_{T+h}(\mathbf{w})-R^*_{T+h}(\mathbf{w})\right| \\ \le\ &\mathbb{E}\left\{\xi^{*-1}_{T+h}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}|L(U_T,\mathbf{w})-L^*(U_T,\mathbf{w})|\right\} \\ =\ &O(\xi^{*-1}_{T+h}(U_T)MqT^{-1/2}l^{-1/2}+\xi^{*-1}_{T+h}(U_T)Mq^2T^{-1}l^{-1}), \end{aligned} \tag{B.6}$$

where the second step uses Condition (C.6), and the last step is due to Condition (C.7).

Next, we observe that

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{SFV}^*(U_T,\mathbf{w})-R^*_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right| \\ \le\ &\xi^{*-1}_{T+h}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{\widetilde{T}l}(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))-\sigma^2_{T+h}-\mathbb{E}L^*(U_T,\mathbf{w})\right| \\ \le\ &\xi^{*-1}_{T+h}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\widetilde{T}^{-1}l^{-1}\left|(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))-(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))\right| \\ &+\xi^{*-1}_{T+h}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\widetilde{T}^{-1}l^{-1}\left|(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))-\mathbb{E}(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))\right| \end{aligned}$$

$$
\begin{aligned}
&+\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\widetilde{T}^{-1}l^{-1}\left|\mathbb{E}(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))-\widetilde{T}l\mathbb{E}\{Y_{T+h}-Y_{T+h}^*(\mathbf{w})\}^2\right| \\
\equiv\ & \xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_1(U_T,\mathbf{w})+\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_2(U_T,\mathbf{w})+\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_3(U_T,\mathbf{w}), \qquad \text{(B.7)}
\end{aligned}
$$

where $\Gamma_1(U_T,\mathbf{w})=(\widetilde{T}l)^{-1}\left|(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))-(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))\right|$,
$\Gamma_2(U_T,\mathbf{w})=(\widetilde{T}l)^{-1}\left|(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))-\mathbb{E}(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))\right|$,
and $\Gamma_3(U_T,\mathbf{w})=(\widetilde{T}l)^{-1}\left|\mathbb{E}(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))-\widetilde{T}l\mathbb{E}\{Y_{T+h}-Y_{T+h}^*(\mathbf{w})\}^2\right|$. Consequently, we need to prove the following equations

$$
\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_1(U_T,\mathbf{w})=o_p(1), \qquad \text{(B.8)}
$$

$$
\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_2(U_T,\mathbf{w})=o_p(1), \qquad \text{(B.9)}
$$

and

$$
\xi_{T+h}^{*-1}(U_T)\sup_{\mathbf{w}\in\mathcal{H}}\Gamma_3(U_T,\mathbf{w})=o_p(1). \qquad \text{(B.10)}
$$

To prove (B.8), we have the decompostion

$$
\begin{aligned}
&(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
=\ & (\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w})+\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w})+\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
=\ & (\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))+(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
&+2(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
=\ & (\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))+(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
&+2(\mathbf{Y}-\boldsymbol{\mu}(\mathbf{w})+\boldsymbol{\mu}(\mathbf{w})-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
=\ & (\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y}-\boldsymbol{\mu}^*(\mathbf{w}))+(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})) \\
&+2(\mathbf{Y}-\boldsymbol{\mu}(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))+2(\boldsymbol{\mu}(\mathbf{w})-\boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\boldsymbol{\mu}^*(\mathbf{w})-\widetilde{\boldsymbol{\mu}}(\mathbf{w})).
\end{aligned}
$$

For any given u_0, we have

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}}\left[\widetilde{\boldsymbol{\mu}}(\mathbf{w})-\boldsymbol{\mu}^*(\mathbf{w})\right]'\mathbf{K}(U_T)\left[\widetilde{\boldsymbol{\mu}}(\mathbf{w})-\boldsymbol{\mu}^*(\mathbf{w})\right] \\
=\ & \sup_{\mathbf{w}\in\mathcal{H}}\left[\mathbf{X}(\widetilde{\boldsymbol{\alpha}}(U_T,\mathbf{w})-\boldsymbol{\alpha}^*(U_T,\mathbf{w}))\right]'\mathbf{K}(U_T)\left[\mathbf{X}(\widetilde{\boldsymbol{\alpha}}(U_T,\mathbf{w})-\boldsymbol{\alpha}^*(U_T,\mathbf{w}))\right]
\end{aligned}
$$

$$= \quad O_p(Tlq) * O_p(MqT^{-1}l^{-1}) = O_p(Mq^2), \tag{B.11}$$

with Lemma 3, Conditions (C.2)-(C.5) and (B.4). Also, with Condition (C.2)-(C.5), it is shown that

$$\sup_{\mathbf{w}\in\mathcal{H}} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}(\mathbf{w}))'\mathbf{K}(U_T)(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| = O_p(M^{1/2}qT^{1/2}l^{1/2}). \tag{B.12}$$

After that, we have

$$\begin{aligned} & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \Gamma_1(U_T, \mathbf{w}) \\ = \quad & O_p(Mq^2\xi_{T+h}^{*-1}(Tl)^{-1} + M^{1/2}q\xi_{T+h}^{*-1}(Tl)^{-1/2}) = o_p(1), \end{aligned} \tag{B.13}$$

where the last step is obtained from Condition (C.7). Therefore, (B.8) is obtained.

To verify (B.9), we have that

$$\begin{aligned} & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \Gamma_2(U_T, \mathbf{w}) \\ = \quad & \xi_{T+h}^{*-1}(U_T)(\widetilde{T}l)^{-1} \sup_{\mathbf{w}\in\mathcal{H}} |\sum_{t=1}^{\widetilde{T}} \sum_{i=1}^{M} \sum_{j=1}^{M} w^i w^j \{k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*} - \mathbb{E} k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*}\}| \\ \leq \quad & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \sum_{i=1}^{M} \sum_{j=1}^{M} w^i w^j |(\widetilde{T}l)^{-1} \sum_{t=1}^{\widetilde{T}} \{k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*} - \mathbb{E} k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*}\}| \\ \leq \quad & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \sum_{i=1}^{M} \sum_{j=1}^{M} |(\widetilde{T}l)^{-1} \sum_{t=1}^{\widetilde{T}} \{k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*} - \mathbb{E} k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*}\}| \\ = \quad & \frac{1}{\xi_{T+h}^{*}(U_T)\sqrt{\widetilde{T}l}} \sum_{i=1}^{M} \sum_{j=1}^{M} \Psi_T(i,j), \end{aligned} \tag{B.14}$$

where $\Psi_T(i,j) = |\frac{1}{\sqrt{\widetilde{T}l}} \sum_{t=1}^{\widetilde{T}} \{k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*} - \mathbb{E} k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*}\}|$. By Theorem 3.49 and 5.20 in White (2014) and Conditions (C.2) and (C.3), we obtain $\Psi_T(i,j) = O_p(1)$ for any i and j. Thus, $\sum_{i=1}^{M} \sum_{j=1}^{M} \Psi_T(i,j) = O_p(M^2)$. Therefore, we have

$$\xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \Gamma_2(U_T, \mathbf{w}) = O_p(\xi_{T+h}^{*-1}(U_T)(Tl)^{-1/2}M^2) = o_p(1), \tag{B.15}$$

where the last step is due to Condition $(C.7)$.

Last, we consider (B.10). We observe that

$$\begin{aligned} & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \Gamma_3(U_T, \mathbf{w}) \\ \leq\ & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} \sum_{i=1}^{M}\sum_{j=1}^{M} w^i w^j \left| \widetilde{T}^{-1} l^{-1} \sum_{t=1}^{\widetilde{T}} \mathbb{E} k_t \epsilon_{t+h}^{(i)*} \epsilon_{t+h}^{(j)*} - \mathbb{E}\epsilon_{T+h}^{(i)*}\epsilon_{T+h}^{(j)*} \right| \\ =\ & o_p(1), \end{aligned} \tag{B.16}$$

where the last step is due to the jointly distributed stationary process $\{U_t, \mathbf{X}_t, Y_{t+h}\}_{t=1}^{\infty}$. Therefore, the proof of Theorem 1 is completed. □

Proof of Theorem 2. Similar to the proof of Theorem 1, it is obtained that

$$\begin{aligned} \mathrm{SFV}(U_T, \mathbf{w}) \ =\ & R_{T+h}^*(\mathbf{w}) + \sigma_{T+h}^2 \\ & + O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}). \end{aligned} \tag{B.17}$$

Denote $\tau(\mathbf{w}) = \sum_{m\in\mathcal{D}} w^m$. Let $\widetilde{\mathbf{w}}$ be a weight vector with $w^m = 0$ for the correctly specified models (i.e., $m \in \mathcal{D}$) and $\widetilde{w}^m = w^m/(1-\tau(\mathbf{w}))$ for all misspecified models (i.e., $m \notin \mathcal{D}$). For any given u_0 and any correctly specified model, it is easy to see that

$$\mu_t^{(m)} - \mu_t = 0 \ \text{ for } m \in \mathcal{D}, \tag{B.18}$$

and

$$\mathbb{E}\{Y_{T+h} - Y_{T+h}^{(m)*}\}^2 = \sigma_{T+h}^2. \tag{B.19}$$

Then, we obtain that

$$\begin{aligned} R_{T+h}^*(\mathbf{w}) \ =\ & \mathbb{E}\left[Y_{T+h} - Y_{T+h}^*(\mathbf{w})\right]^2 - \sigma_{T+h}^2 \\ =\ & \mathbb{E}\left[\sum_{m=1}^{M} w^m \{Y_{T+h} - \epsilon_{T+h} - Y_{T+h}^{(m)*}\}\right]^2 \\ =\ & (1-\tau(\mathbf{w}))^2 \mathbb{E}\left[\sum_{m\notin\mathcal{D}} (1-\tau(\mathbf{w}))^{-1} w^m \{Y_{T+h} - \epsilon_{T+h} - Y_{T+h}^{(m)*}\}\right]^2 \\ =\ & (1-\tau(\mathbf{w}))^2 R_{T+h}^*(\widetilde{\mathbf{w}}). \end{aligned} \tag{B.20}$$

By replacing $\mathbf{w}$ with its estimator and based on (B.17) and (B.20), we have

$$\mathrm{SFV}^*(U_T, \widehat{\mathbf{w}}(U_T)) = (1-\tau(\widehat{\mathbf{w}}(U_T)))^2 R^*_{T+h}(\widehat{\widetilde{\mathbf{w}}}(U_T)) + O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}).$$

Let $\boldsymbol{\lambda}$ be a weight vector with $\sum_{m\in\mathcal{D}} \lambda^m = 1$. Also, we obtain $R^*_{T+h}(\boldsymbol{\lambda}) = 0$ based on (B.18). Then, it is shown that

$$\mathrm{SFV}^*(u_0, \boldsymbol{\lambda}) = O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}),$$

and

$$\begin{aligned} & (1-\tau(\widehat{\mathbf{w}}(U_T)))^2 R^*_{T+h}(\widehat{\widetilde{\mathbf{w}}}(U_T)) + O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}) \\ \leq\ & \mathrm{SFV}^*(u_0, \boldsymbol{\lambda}) = O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}). \end{aligned} \tag{B.21}$$

based on the fact that $\widehat{\mathbf{w}}(U_T) = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathrm{SFV}^*(U_T, \mathbf{w})$. Therefore, we derive that

$$\begin{aligned} & (1-\tau(\widehat{\mathbf{w}}(U_T)))^2 \inf_{\mathbf{w}\in\widetilde{\mathcal{H}}} R^*_{T+h}(\mathbf{w}) + O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}) \\ \leq\ & O_p(qM^{1/2}T^{-1/2}l^{-1/2} + M^2T^{-1/2}l^{-1/2} + Mq^2(Tl)^{-1}), \end{aligned} \tag{B.22}$$

where $\tau(\widehat{\mathbf{w}}(U_T))$ is the estimator of $\tau(\mathbf{w})$ for any given U_T, i.e., $\tau(\widehat{\mathbf{w}}(U_T)) = \sum_{j\in\mathcal{D}} \widehat{w}^j(U_T)$. Combined with Condition (C.8), for any given u_0 we obtain $\tau(\widehat{\mathbf{w}}(U_T)) \xrightarrow{p} 1$, and thus, the proof of Theorem 2 is completed. □

Proof of Corollary 1. When all candidate models are misspecified, the results of Theorem 1 guarantee that our proposed procedure achieves the minimum prediction risk among all possible estimators. Consequently, the risk associated with our approach cannot exceed the risk of simple model averaging. We now turn our attention to the scenario where the correctly specified models are encompassed within the set of candidate models. In this scenario, the simple average procedure ensures that

$$\begin{aligned} \widehat{Y}_{T+h}(\mathbf{1}/M) &= \sum_{m=1}^{M} \frac{1}{M} \widehat{Y}^{(m)}_{T+h} = \sum_{m=1}^{M} \frac{1}{M} \mathbf{X}_T' \left[\boldsymbol{\Pi}^{(m)\prime} \widehat{\boldsymbol{\alpha}}^{(m)}(U_T) - \boldsymbol{\alpha}(U_T) \right] + \mu_T, \\ &= \mu_T + \sum_{m=1}^{M} \frac{1}{M} \mathbf{X}_T' \left[\boldsymbol{\Pi}^{(m)\prime} \left(\widehat{\boldsymbol{\alpha}}^{(m)}(U_T) - \boldsymbol{\alpha}^{(m)*}(U_T) + \boldsymbol{\alpha}^{(m)*}(U_T) \right) - \boldsymbol{\alpha}(U_T) \right], \end{aligned}$$

$$
\begin{aligned}
&= \mu_T + \sum_{m=1}^{M} \frac{1}{M} \mathbf{X}_T' \left[\mathbf{\Pi}^{(m)'} \left(\widehat{\boldsymbol{\alpha}}^{(m)}(U_T) - \boldsymbol{\alpha}^{(m)*}(U_T) \right) \right] + \sum_{m=1}^{M} \frac{1}{M} \mathbf{X}_T' \left[\mathbf{\Pi}^{(m)'} \boldsymbol{\alpha}^{(m)*}(U_T) - \boldsymbol{\alpha}(U_T) \right], \\
&= \mu_T + O_p(T^{-1/2} l^{-1/2} M^{1/2} q) + v_q, \qquad \text{(B.23)}
\end{aligned}
$$

given Lemma 3, where v_q has the same order as KM^{-1}. Thus, the prediction risk is $R_{T+h}(\mathbf{1}/M) = O_p(T^{-1}l^{-1}Mq^2) + v_q^2$. Our SVMA prediction has

$$
\begin{aligned}
&\widehat{Y}_{T+h}(\widehat{\mathbf{w}}(U_T)) = \sum_{m=1}^{M} \widehat{w}^m(U_T) \widehat{Y}_{T+h}^{(m)} \\
&= \mu_T + \sum_{m \in \mathcal{D}} \widehat{w}^m(U_T) \left[\widehat{Y}_{T+h}^{(m)} - \mu_T \right] + \sum_{m \notin \mathcal{D}} \widehat{w}^m(U_T) \left[\widehat{Y}_{T+h}^{(m)} - \mu_T \right] \\
&= \mu_T + O_p(T^{-1/2} l^{-1/2} M^{1/2} q) + O_p(1 - \widehat{\tau}(\mathbf{w}(U_T))) \\
&= \mu_T + O_p(T^{-1/2} l^{-1/2} M^{1/2} q) \\
&\quad + O_p\left[T^{-1/4} l^{-1/4} (q^{1/2} M^{1/4} + M + M^{1/2} q T^{-1/4} l^{-1/4}) \left\{ \inf_{\mathbf{w} \in \widetilde{\mathcal{H}}} (\mathbb{E} L_T^*(U_T, \mathbf{w}))^{-1/2} \right\} \right] \\
&= \mu_T + O_p(T^{-1/2} l^{-1/2} M^{1/2} q) + O_p(T^{-1/4} l^{-1/4} (q^{1/2} M^{1/4} + M)) \qquad \text{(B.24)}
\end{aligned}
$$

Hence, the prediction risk is $R_{T+h}(\widehat{\mathbf{w}}(U_T)) = O_p(T^{-1}l^{-1}Mq^2 + T^{-1/2}l^{-1/2}(qM^{1/2} + M^2))$, which is smaller than $R_{T+h}(\mathbf{1}/M)$ as long as $T^{-1}l^{-1}Mq^2/v_q^2 = o(1)$ and $T^{-1/2}l^{-1/2}(qM^{1/2} + M^2)/v_q^2 = o(1)$. This implies that $T^{-1}l^{-1}M^3q^2K^{-2} = o(1)$ and $T^{-1/2}l^{-1/2}K^{-2}(qM^{5/2} + M^4) = o(1)$. □

B.2 SVMA for Ultra-High Dimensional Setting

Proof of Corollary 2. The proof of Corollary 2 is an improvement of the proof of Theorem 3 in Zhang et al. (2016). To verify Corollary 2, it is equivalent to show that for any given u_0 and for any $\delta > 0$,

$$
\Pr\left\{ \left| \frac{\inf_{\mathbf{w} \in \mathcal{H}} R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))} - 1 \right| > \delta \right\} \xrightarrow{P} 0. \qquad \text{(B.25)}
$$

First, we define $\mathrm{DF}(U_T, \mathbf{w}) = \mathrm{SFV}^*(U_T, \mathbf{w}) - R_{T+h}(\mathbf{w})$. Based on (B.3),

$$
\sup_{\mathbf{w} \in \mathcal{H}} |\mathrm{DF}(U_T, \mathbf{w}) / R_{T+h}^*(\mathbf{w})| = o_p(1)
$$

for any given u_0. With Conditions (C.7) and (C.10), it is observed that for any given u_0

$$\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{v_T(U_T)}{R^*_{T+h}(\mathbf{w})}\right| = o(1),$$

and

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})}\right| &= \left\{\inf_{\mathbf{w}\in\mathcal{H}}\left|\frac{R_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|\right\}^{-1} \\ &\leq \left\{1-\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R_{T+h}(\mathbf{w})-R^*_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|\right\}^{-1} \to 1,\end{aligned}$$

where the last step is obtained from (B.2). Similarly, we have

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})-v_T(U_T)}\right| \\ \leq\ &\left\{1-\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R_{T+h}(\mathbf{w})-R^*_{T+h}(\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|-\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{v_T(U_T)}{R^*_{T+h}(\mathbf{w})}\right|\right\}^{-1} \to 1.\end{aligned}$$

Finally, we prove (B.25). It is observed that

$$\begin{aligned}&\Pr\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}-1\right|>\delta\right\} \\ =\ &\Pr\left\{\left|\frac{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))-\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}\right|>\delta\right\} \\ =\ &\Pr\left\{\left|\frac{\mathrm{SFV}^*(U_T,\widehat{\mathbf{w}}^*(U_T))-\mathrm{DF}(U_T,\widehat{\mathbf{w}}^*(U_T))-\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}\right|>\delta\right\} \\ =\ &\Pr\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{H}^*}[R_{T+h}(\mathbf{w})+\mathrm{DF}(U_T,\mathbf{w})]-\mathrm{DF}(U_T,\widehat{\mathbf{w}}^*(U_T))-\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}\right|>\delta\right\} \\ \leq\ &\Pr\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{H}^*}[R_{T+h}(\mathbf{w})+\mathrm{DF}(U_T,\mathbf{w})]-\mathrm{DF}(U_T,\widehat{\mathbf{w}}^*(U_T))-\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}\right|>\delta|\mathbf{w}_T\in\mathcal{H}^*\right\} \\ &\times\Pr(\mathbf{w}_T\in\mathcal{H}^*)+\Pr(\mathbf{w}_T\notin\mathcal{H}^*) \\ \leq\ &\Pr\left\{\left|\frac{\mathrm{DF}(U_T,\mathbf{w}_T)-\mathrm{DF}(U_T,\widehat{\mathbf{w}}^*(U_T))+v_T(U_T)}{R_{T+h}(\widehat{\mathbf{w}}^*(U_T))}\right|>\delta\right\}+\Pr(\mathbf{w}_T\notin\mathcal{H}^*) \\ \leq\ &\Pr\left\{\left|\frac{\mathrm{DF}(U_T,\mathbf{w}_T)}{\inf_{\mathbf{w}\in\mathcal{H}}R_{T+h}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{DF}(U_T,\mathbf{w})}{R_{T+h}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{v_T(U_T)}{R_{T+h}(\mathbf{w})}\right|>\delta\right\}+\Pr(\mathbf{w}_T\notin\mathcal{H}^*) \\ \leq\ &\Pr\left\{\left|\frac{\mathrm{DF}(U_T,\mathbf{w}_T)}{R_{T+h}(\mathbf{w}_T)-v_T(U_T)}\right|+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{DF}(U_T,\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})}\right|\right. \\ &\left.+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{v_T(U_T)}{R^*_{T+h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})}\right|>\delta\right\}+\Pr(\mathbf{w}_T\notin\mathcal{H}^*) \\ \leq\ &\Pr\left\{\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{DF}(U_T,\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})-v_T(U_T)}\right|+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{\mathrm{DF}(U_T,\mathbf{w})}{R^*_{T+h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})}\right|\right.\end{aligned}$$

$$+\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{v_T(U_T)}{R^*_{T+h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{R^*_{T+h}(\mathbf{w})}{R_{T+h}(\mathbf{w})}\right|>\delta\Bigg\}+\Pr(\mathbf{w}_T\notin\mathcal{H}^*)\xrightarrow{P}0.$$

Thus, (B.25) is obtained and therefore, the proof of Corollary 2 is completed. □

B.3 SVMA for State-Varying Factor Models

For any given u_0, the notations like $\widehat{\boldsymbol{\mu}}^{(m)}$ and $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ in Section 2 are modified by replacing $\mathbf{X}_t^{(m)}$ with $\widehat{\mathbb{Z}}_t^{(m)}$ and $\boldsymbol{\alpha}^{(m)}(u_0)$ with $\boldsymbol{\Psi}^{(m)}(u_0)$, respectively. The model averaging estimator for $\boldsymbol{\Psi}(u_0)$ is given by

$$\widehat{\boldsymbol{\Psi}}(u_0,\mathbf{w})=\sum_{m=1}^{M}w^m\boldsymbol{\Pi}^{(m)\prime}\widehat{\boldsymbol{\Psi}}^{(m)}(u_0),\tag{B.26}$$

where $\boldsymbol{\Pi}^{(m)}$ is replaced by a selected matrix of size $(q_m+r\times r_m)\times(q+r\times r_0)$ mapping $\mathbb{Z}_t$ to $\mathbb{Z}_t^{(m)}$, and $\mathbb{Z}_t$ is the predictor vector including all predictors from candidate models. On the other hand, the forward-validation estimation-based notations like $\widetilde{\boldsymbol{\mu}}^{(m)}$, and $\widetilde{\boldsymbol{\mu}}(\mathbf{w})$ in Section 2 are modified by replacing $\mathbf{X}_t^{(m)}$ with $\widehat{\mathbb{Z}}_t^{(m)}$, and $\widehat{\boldsymbol{\alpha}}_{[-t]}^{(m)}(u_0)$ with $\widehat{\boldsymbol{\Psi}}_{[-t]}^{(m)}(u_0)$, the parameter estimator after removing observations; see the following detailed notations.

Replacing $\mathbf{X}_t$ with $\widehat{\mathbb{Z}}_t$, define $\widehat{\mathbb{Z}}^{(m)}$ as a $(T-h)\times(q_m+r\times r_m)$ matrix with rows $\widehat{\mathbb{Z}}_t^{(m)\prime}$. We obtain $\widehat{\mathbb{Z}}_{[t]}^{(m)}=\boldsymbol{\phi}_t\widehat{\mathbb{Z}}^{(m)}$, the sets to be removed, and $\widehat{\mathbb{Z}}_{[-t]}^{(m)}$ as the remaining sets of $\mathbb{Z}^{(m)}$ after removing $\mathbb{Z}_{[t]}^{(m)}$. For any given u_0, the following local constant estimator $\widehat{\boldsymbol{\Psi}}_{[-t]}^{(m)}(u_0)$ is obtained as

$$\widehat{\boldsymbol{\Psi}}_{[-t]}^{(m)}(u_0)=\left(\widehat{\mathbb{Z}}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\widehat{\mathbb{Z}}_{[-t]}^{(m)}\right)^{-1}\widehat{\mathbb{Z}}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\mathbf{Y}_{[-(t+h)]}.\tag{B.27}$$

And the leave-h-out forward-validation estimator $\widetilde{\mu}_t^{(m)}(u_0)$ of $\mu_t^{(m)}$ is

$$\widetilde{\mu}_t^{(m)}(u_0)=\boldsymbol{\pi}_t\widehat{\mathbb{Z}}_{[t]}^{(m)}\widehat{\boldsymbol{\Psi}}_{[-t]}^{(m)}(u_0).\tag{B.28}$$

Thus, the forward validation averaging estimator of μ_t is $\widetilde{\mu}_t(\mathbf{w})=\sum_{m=1}^{M}w^m\widetilde{\mu}_t^{(m)}(u_0)$ and then $\widetilde{\boldsymbol{\mu}}(\mathbf{w})=(\widetilde{\mu}_1(\mathbf{w}),\ldots,\widetilde{\mu}_{T-h}(\mathbf{w}))'$.

Within $\widehat{\mathbb{Z}}_t$ and $\widehat{\boldsymbol{\Psi}}_{[-t]}^{(m)}(u_0)$, the state-dependent forward-validation criterion is

$$\mathrm{SFV}(u_0,\mathbf{w})=\frac{1}{(T-h)l}(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(u_0)(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}(\mathbf{w})).\tag{B.29}$$

For any given u_0, minimizing $\text{SFV}(u_0, \mathbf{w})$ with respect to $\mathbf{w}$, we have

$$\widehat{\mathbf{w}}(u_0) = \arg\min_{\mathbf{w}\in\mathcal{H}} \text{SFV}(u_0, \mathbf{w}).$$

Then, the SVMA prediction for Y_{T+h} is $\widehat{Y}_{T+h}(\widehat{\mathbf{w}}(U_T)) = \sum_{m=1}^{M} \widehat{w}^m(U_T)\widehat{Y}_{T+h}^{(m)}$, where $\widehat{\mathbf{w}}(U_T) = (\widehat{w}^1(U_T), \ldots, \widehat{w}^M(U_T))'$.

Condition (C.11). *For the m-th candiate model and given any point u_0, there exists a limit $\mathbf{\Psi}^{(m)*}(u_0)$ such that $||\widehat{\mathbf{\Psi}}^{(m)}(u_0) - \mathbf{\Psi}^{(m)*}(u_0)|| = O_p(q^{1/2}T^{-1/2}l^{-1/2})$, as $T, N \to \infty$.*

Condition (C.12). *For each i, $\{\mathbf{X}_t, \mathbf{U}_t, \mathbf{V}_{it}, \mathbf{F}_t, Y_{t+h}\}$ is α-mixing process of size $-\tau/(\tau-2)$ with $\tau > 2$, $\sup_{1\le t\le T-h} ||\mathbb{Z}_t^{(m)}||/\sqrt{q} = O_p(1)$ uniformly for all m, and for some positive constant C_0, $\zeta_{\min}((T-h)^{-1}l^{-1}\mathbb{Z}^{(m)'}\mathbf{K}(u_0)\mathbb{Z}^{(m)}) \ge C_0$.*

Remark 7. *Condition (C.11) is a high-level assumption that ensures the estimator $\widehat{\mathbf{\Psi}}^{(m)}$ of $\mathbf{\Psi}^{(m)}$ in each candidate model converges to a well-defined limit $\mathbf{\Psi}^{(m)*}$, and it concerns the convergence rate of local constant estimation in potentially misspecified models. The limit $\mathbf{\Psi}^{(m)*}$ can be interpreted as a pseudo-true value. Condition (C.11) is widely employed in analyzing the asymptotic properties of model averaging estimators, e.g., Assumption 1 of Zhang and Zhang (2023) and equation (6) of Zhang et al. (2016). Without the estimated factors, Condition (C.11) is essentially equivalent to Lemma 3. If the parameters are constant and without factors, Condition (C.11) holds under appropriate primitive conditions; see Assumptions in Theorem 3.2 of White (1982). Besides, if the m-th candidate model is correctly specified, the estimated common factor is a consistent estimator for the latent factor $\mathbf{F}_t$ up to some state-dependent rotation matrix $\mathbf{H}(u_0)$ under some primitive conditions for any given u_0; see Theorem 1 of Pelger and Xiong (2022)*[5]*. After that, Condition (C.11)*

[5]Condition (C.11) implies certain regularity conditions for factor models when the factor-augmented regression incorporates the generated factors obtained from the first-step estimation as predictors. For instance, the fourth moment of the factor is bounded both with and without the state filtration, and the conditional covariance matrix of the factors needs to be positive definite, which is the state-conditional version of Assumption A in Bai (2003) and consistent with Assumption 3 of Pelger and Xiong (2022). Furthermore, factor loadings are deterministic and Lipschitz continuous with respect to the state, as specified in Assumption 4 of Pelger and Xiong (2022).

could be obtained similarly to Theorem 3.1 of Chen et al. (2024), which allows both q and r to diverge at a slower rate than the sample size $\widetilde{T}l$. Condition (C.12) is an extension of Condition (C.2).

Proof of Corollary 3. Based on Condition (C.11), for any given u_0, it is observed that

$$\max_{1\leq m\leq M} ||\widehat{\mathbf{\Psi}}^{(m)}(u_0) - \mathbf{\Psi}^{(m)*}(u_0)|| = O_p(M^{1/2}q^{1/2}T^{-1/2}l^{-1/2}).$$

Then, we have

$$||\widehat{\mathbf{\Psi}}(U_T, \mathbf{w}) - \mathbf{\Psi}^*(U_T, \mathbf{w})|| = O_p(M^{1/2}q^{1/2}T^{-1/2}l^{-1/2}). \tag{B.30}$$

Similar to the proof Theorem 1, we have that

$$\sup_{\mathbf{w}\in\mathcal{H}} |L(U_T, \mathbf{w}) - L^*(U_T, \mathbf{w})| = O_p(MqT^{-1/2}l^{-1/2} + Mq^2T^{-1}l^{-1}). \tag{B.31}$$

For any given u_0, we have

$$\sup_{\mathbf{w}\in\mathcal{H}} [\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})]' \mathbf{K}(U_T) [\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})] = O_p(Mq^2), \tag{B.32}$$

with Lemma 3, Conditions (C.2)-(C.5) and (B.4). Also, with Condition (C.2)-(C.5), it is shown that

$$\sup_{\mathbf{w}\in\mathcal{H}} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}(\mathbf{w}))'\mathbf{K}(U_T)(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| = O_p(M^{1/2}qT^{1/2}l^{1/2}). \tag{B.33}$$

After that, we have

$$\begin{aligned} & \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} (\widetilde{T}l)^{-1} |(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})) - (\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w}))| \\ = \; & O_p(Mq^2\xi_{T+h}^{*-1}(Tl)^{-1} + M^{1/2}q\xi_{T+h}^{*-1}(Tl)^{-1/2}) = o_p(1), \end{aligned} \tag{B.34}$$

where the last step is obtained from Condition (C.7). Similar to the proof of (B.9), with Conditions (C.3) and (C.12), we have

$$\begin{aligned} \xi_{T+h}^{*-1}(U_T) \sup_{\mathbf{w}\in\mathcal{H}} (\widetilde{T}l)^{-1} & |(\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w})) \\ & -\mathbb{E}(\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(U_T)(\mathbf{Y} - \boldsymbol{\mu}^*(\mathbf{w}))| = o_p(1), \end{aligned}$$

based on Theorem 5.20 of White (2014). Therefore, the proof of Corollary 3 is completed. □

Appendix C

The following two tables present the complete numerical results of the relative MSFE gains for SVMA, SA, SAICc, SAIC and SBIC over AICc method discussed in Section 5. Note that the first column in both tables indicates the initial time point for the relative MSFE calculation, the final time point is December 2020. The relative MSFE gain is computed as:

$$\text{MSFE}^{h}_{(i)}\left(T_1, T_2\right) = 1 - \frac{\sum_{t=T_1}^{T_2} \widehat{\epsilon}^{2}_{t,(i)}}{\sum_{t=T_1}^{T_2} \widehat{\epsilon}^{2}_{t,(\text{AIC}_c)}} \quad \text{with} \quad T_2 = 2020:12,$$

where the forecast data $T_1 \in [2019:1, 2020:1]$, and i represents the SVMA, SA, SAICc, SAIC, and SBIC method, respectively.

Table C.1: Relative MSFE values for stock return forecasting.

$h=1$	SVMA	SA	SAICc	SAIC	SBIC
Jan. 2019	0.1613	-0.1038	-0.0936	-1.2645	0.1405
Feb. 2019	0.1639	-0.1023	-0.0921	-1.2629	0.1417
Mar. 2019	0.1664	-0.1037	-0.0933	-1.2942	0.1474
Apl. 2019	0.1693	-0.0931	-0.0831	-1.2865	0.1488
May. 2019	0.1697	-0.0939	-0.0839	-1.2831	0.1491
Jun. 2019	0.1765	-0.0871	-0.0774	-1.2831	0.1518
Jul. 2019	0.2225	-0.0836	-0.0738	-1.2845	0.1516
Aug. 2019	0.2226	-0.0832	-0.0734	-1.2822	0.1517
Sep. 2019	0.2232	-0.0833	-0.0735	-1.2789	0.1520
Oct. 2019	0.2233	-0.0835	-0.0737	-1.2797	0.1519
Nov. 2019	0.2092	-0.0803	-0.0698	-1.3110	0.1500
Dec. 2019	0.2088	-0.0768	-0.0665	-1.3174	0.1481
Jan. 2020	0.2132	-0.0796	-0.0692	-1.3449	0.1494
$h=2$					
Jan. 2019	0.5387	-1.7699	-1.9438	-0.5429	0.4972
Feb. 2019	0.5432	-1.7839	-1.9592	-0.5429	0.5028
Mar. 2019	0.5455	-1.7926	-1.9686	-0.5477	0.5054
Apl. 2019	0.5455	-1.7924	-1.9684	-0.5477	0.5055
May. 2019	0.5448	-1.7960	-1.9722	-0.5492	0.5066
Jun. 2019	0.5450	-1.7971	-1.9734	-0.5498	0.5068
Jul. 2019	0.5543	-1.7880	-1.9654	-0.5490	0.5067
Aug. 2019	0.5494	-1.8296	-2.0095	-0.5719	0.5018
Sep. 2019	0.5497	-1.8301	-2.0101	-0.5709	0.5024
Oct. 2019	0.5501	-1.8312	-2.0113	-0.5719	0.5026
Nov. 2019	0.5565	-1.8464	-2.0281	-0.5772	0.5095
Dec. 2019	0.5558	-1.8521	-2.0342	-0.5780	0.5107
Jan. 2020	0.5708	-1.8679	-2.0532	-0.5769	0.5176

Table C.2: Relative MSFE values for macroeconomic forecasting.

RPI	SVMA	SA	SAICc	SAIC	SBIC
Jan. 2022	0.8373	0.4767	0.4620	-50.6540	-0.9070
Feb. 2022	0.8336	0.4216	0.4053	-33.8541	-1.6214
Mar. 2022	0.8308	0.4091	0.3926	-35.5296	-1.8095
Apl. 2022	0.8343	0.4112	0.3947	-35.6883	-1.8224
May. 2022	0.8046	0.2696	0.2503	-43.8053	-2.5175
Jun. 2022	0.8062	0.2697	0.2503	-42.0408	-2.5160
Jul. 2022	0.8324	0.4078	0.3899	-5.8861	-1.8910
Aug. 2022	0.8315	0.3742	0.3551	-5.7306	-2.0329
Sep. 2022	0.8989	0.4150	0.3950	-5.5265	-2.3110
Oct. 2022	0.9005	0.4143	0.3947	-4.8809	-2.4465
Nov. 2022	0.8712	0.2477	0.2249	-6.4299	-3.5172
Dec. 2022	0.8710	0.2430	0.2205	-6.5930	-3.6560
Jan. 2023	0.8736	0.2394	0.2160	-5.9873	-3.7624
TB3MS					
Jan. 2022	0.4641	0.1259	0.1264	-7.6447	-0.4958
Feb. 2022	0.4755	0.1392	0.1394	-12.5431	-1.0309
Mar. 2022	0.5586	0.0829	0.0850	-19.1140	-1.7566
Apl. 2022	0.5255	0.0316	0.0343	-21.0569	-2.1595
May. 2022	0.7067	0.2450	0.2471	-20.2310	-2.0256
Jun. 2022	0.7041	0.2325	0.2347	-20.6188	-1.7957
Jul. 2022	0.6080	0.1247	0.1286	-32.7638	-3.5055
Aug. 2022	0.6076	0.0851	0.0893	-29.3800	-3.5903
Sep. 2022	0.7382	0.0152	0.0245	-42.7100	-5.2612
Oct. 2022	0.7447	0.0055	0.0149	-18.9720	-3.0062
Nov. 2022	0.7711	-0.0031	0.0063	-19.0943	-2.9949
Dec. 2022	0.8566	0.1018	0.1098	-5.9093	-3.1701
Jan. 2023	0.9041	0.3444	0.3497	-9.4143	-5.6199
PAYEMS					
Jan. 2022	0.5174	0.2773	0.2560	-43.5201	-43.3931
Feb. 2022	0.4356	0.2805	0.2517	-38.2225	-37.9855
Mar. 2022	0.5928	0.4487	0.4226	-48.4543	-48.1165
Apl. 2022	0.6378	0.4843	0.4568	-13.5151	-13.1747
May. 2022	0.6066	0.4861	0.4597	-2.5746	-2.2582
Jun. 2022	0.6121	0.4876	0.4617	-2.4950	-2.1786
Jul. 2022	0.5321	0.4442	0.4031	-2.4750	-2.0453
Aug. 2022	0.6587	0.5089	0.4678	-2.6533	-2.2018
Sep. 2022	0.6777	0.4971	0.4377	-4.5569	-3.8102
Oct. 2022	0.6638	0.5319	0.4764	-4.7950	-4.0083
Nov. 2022	0.7078	0.4855	0.4295	-4.9353	-4.0708
Dec. 2022	0.6112	0.4220	0.3663	-6.6503	-5.4877
Jan. 2023	0.5365	0.4965	0.4468	-6.8520	-5.3705